UNIVERSITY OF CALIFORNIA SAN DIEGO

# Exploring the Adaptive Capabilities of *Escherichia coli*: Perturbations from a Typical Life Cycle and Resulting Evolutionary Compensations

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

in

Bioengineering

by

Troy Evan Sandberg

Committee in charge:

Bernhard Ø. Palsson, Chair
Lin Chao
Adam M. Feist
Christian M. Metallo
Justin Meyer
Scott Rifkin

2018

The dissertation of Troy Evan Sandberg is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

DEDICATION

For my parents.

# EPIGRAPH

*Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.*

—Richard Feynman

*Nothing in Biology makes sense except in the light of evolution.*

— Theodosius Dobzhansky

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to my friends and family for love and support. Thanks to Bernhard Palsson, Adam Feist, Ryan LaCroix, Richard Szubin, Connor Olson, Patrick Phaneuf, and all other collaborators and members of the Palsson lab. Thanks to the National Science Foundation for selecting me for a Graduate Research Fellowship. Thanks to the Siebel Scholar Foundation for selecting me as a Siebel Scholar.

Chapter 1 borrows in part from a manuscript in preparation for submission: Sandberg TE, Weng LL, Salazar MJ, Palsson BO, Feist AM. “Adaptive Laboratory Evolution: A Valuable Technique for Discovery and Industrial Biotechnology” The dissertation author is one of the authors of the review.

Chapter 2 is a reprint of a published manuscript: Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM. 2014. "Evolution of *Escherichia coli* to 42 °C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations." *Molecular Biology and Evolution*, 31(10):2647-62. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 3 is a reprint of a published manuscript: 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(3):e0151130. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 4 is a reprint of a published manuscript: Sandberg TE, Lloyd CJ, Palsson BO, Feist AM. 2017. “Laboratory Evolution to Alternating Substrate Environments

Yields Distinct Phenotypic and Genetic Adaptive Strategies." *Applied and Environmental Microbiology*, 83(13):e00410-17. The dissertation author was the primary author of the paper and was responsible for the research.

Chapter 5 is a version of a manuscript in preparation for submission: Sandberg TE, Catoiu E, Szubin R, Phaneuf PV, Feist AM, Palsson BO. "Synthetic Cross-Domain Gene Replacement and Evolutionary Assimilation of Major Glycolytic Enzymes into *E. coli*." The dissertation author was the primary author of the paper and was responsible for the research.

## VITA

2012 B. S. in Bioengineering, California Institute of Technology

2018 Ph. D. in Bioengineering, University of California San Diego

## PUBLICATIONS

**Sandberg TE**, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM. 2014. "Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations." *Molecular Biology and Evolution*, 31(10):2647-62

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." *Applied and Environmental Microbiology*, 81.1, pp. 17–30.

**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(3):e0151130

**Sandberg TE**, Lloyd CJ, Palsson BO, Feist AM. 2017. "Laboratory evolution to alternating substrate environments yields distinct phenotypic and genetic adaptive strategies." *Applied and Environmental Microbiology*, 83(13):e00410-17

McCloskey D, Xu S, **Sandberg TE**, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. 2018. "Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism" *Nature Communications*, 9:3796

McCloskey D, Xu S, **Sandberg TE**, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. 2018. "Adaptation to the coupling of glycolysis to toxic methylglyoxal production in tpiA deletion strains of Escherichia coli requires synchronized and counterintuitive genetic changes." *Metabolic Engineering*, 48:82-93

McCloskey D, Xu S, **Sandberg TE**, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. 2018. "Growth adaptation of gnd and sdhCB Escherichia coli deletion strains diverges from a similar initial perturbation of the transcriptome." *Frontiers in Microbiology*, 9:1793

McCloskey D, Xu S, **Sandberg TE**, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. 2018. "Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System." *Metabolic Engineering*, 48:233-242

McCloskey D, Xu S, **Sandberg TE**, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. 2018. “Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in Escherichia coli pgi knockout evolutions.” *Applied and Environmental Microbiol*ogy, pii: AEM.00823-18

Lloyd CJ, King Z, **Sandberg TE**, Hefner Y, Olson C, Phaneuf P, O’Brien E, Feist AM. “Model-driven design and evolution of non-trivial synthetic syntrophic pairs” Manuscript under review

Anand A, Yang L, Sastry AV, Olson CA, Catoiu E, Phaneuf PV, **Sandberg TE**, Xu S, Hefner Y, Szubin R, Feist AM, Palsson BO. “Selection pressure applied in experimental evolution leads to pseudogene repair” Manuscript under review

Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, **Sandberg TE**, Seo SW, Kim D, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Szubin R, Hefner Y, Feist AM, Palsson BO. “Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms” Manuscript under review

Guzmán GI, **Sandberg TE**, LaCroix RA, Nyerges A, Papp H, de Raad M, King ZA, Northen TR, Notebaart RA, Pál C, Palsson BO, Papp B, Feist AM. "Enzyme promiscuity shapes evolutionary innovation and optimization." Manuscript under review.

Chen K, Olson C, Anand A, **Sandberg TE**, Gao Y, Mih N, Palsson BO. “Bacterial fitness landscapes stratify based on discrete metabolic and proteome states.” Manuscript under review

Cheng C, O’Brien EJ, McCloskey D, Utrilla J, Olson C, LaCroix RA, **Sandberg TE**, Feist AM, Palsson BO, King ZA. “Laboratory evolution reveals a two-dimensional rate-yield tradeoff in microbial metabolism.” Manuscript under review

ABSTRACT OF THE DISSERTATION

**Exploring the Adaptive Capabilities of *Escherichia coli*: Perturbations from a Typical Life Cycle and Resulting Evolutionary Compensations**

by

Troy Evan Sandberg

Doctor of Philosophy in Bioengineering

University of California San Diego, 2018

Professor Bernhard Ø. Palsson, Chair

Evolution underlies the entirety of Earth's biodiversity, as all species diverged from the last universal common ancestor billions of years ago. Although able to effect incredible changes over long periods, the need for multiple generations of mutation and competition renders evolution nearly imperceptible, at the timescale of human observation, for all but the most quickly reproducing organisms. Thus microbial adaptation, given microbes' rapid generation time and enormous population sizes, is perhaps most pressing to understand. This unavoidable evolutionary process facilitates the rise and spread of antibiotic resistance, and frequently countervails attempts to genetically engineer organisms for human purposes.

The bacterium *Escherichia coli*, easily the most highly studied microbe to date, is an ideal model by which to investigate evolution. With both clinical and biotechnological relevance, a thorough understanding of the adaptive principles governing *E. coli*

evolution is of great importance. In this dissertation I seek to probe the adaptive capabilities of *E. coli* using custom robotics systems that function as 'evolution machines.' Enabled by this automation, adaptive walks along the fitness landscape can be tracked in real-time with experimental throughput, data quality, and environmental control impossible to replicate manually.

I subject *E. coli* to stressful perturbations and analyze the mechanisms by which it evolves to restore robust growth, using data types such as phenotypic characterization, whole genome sequencing, and transcriptomics. I demonstrate the remarkable adaptive flexibility of *E. coli* as it readily evolves to tolerate elevated temperatures, altered isotopic composition, rapidly fluctuating growth environments, and even replacement of important native genes with foreign DNA. Overall, these studies establish condition-specific evolutionary responses, general mechanisms for growth rate improvement, and guiding principles for the successful use of laboratory evolution experiments as a tool for biological discovery and engineering.

# Chapter 1

# Introduction

The bacterium *Escherichia coli* is likely the most highly studied and well-characterized model organism to date. While an essential component of the human microbiome (and occasional pathogen), *E. coli*'s rapid growth and ease of culturability established it as a laboratory workhorse more than half a century ago. Technology advancements since then have witnessed the development of powerful techniques for genetic manipulation of *E. coli*, making it not just biologically and scientifically important but also commercially – strains are frequently engineered to produce valuable proteins and chemicals [1]. Inherent to its short duplication time, *E. coli* is also quick to evolve, which can be problematic biologically (pathogenic strain evolution), scientifically (unwanted mutation development over time), and commercially (mutations that decrease product titer). Understanding the *E. coli* adaptation process is therefore of great importance.

## 1.1 Adaptive Laboratory Evolution

Adaptive laboratory evolution, or ALE, is increasingly being used as a powerful tool for both biological discovery and engineering [2]. In their simplest form, ALE experiments consist of prolonged culturing of cells in a chosen environment to select for

spontaneous mutants able to outcompete the ancestor (Figure 1.1). ALE works robustly in microbes due to the ease with which large populations of rapidly dividing cells can be maintained; typical mutation rates and genome sizes ensure extensive sampling of the adaptive space, providing ample genetic diversity from which beneficial mutants will be naturally enriched [3]. Furthermore, the rise in the use of ALE has been fueled significantly by low cost, high throughput DNA sequencing, which when paired with appropriate bioinformatics tools allows facile identification of causal mutations. By relying on natural selection to enrich for mutants with increased fitness, phenotypic improvements are realized without requiring *a priori* knowledge of the genetic alterations necessary to effect such changes. In this way ALE can both optimize strain properties and provide insight into selective features governing the adaptation process.

Over the course of my graduate career I have been deeply involved in the development and optimization of "ALE machine" technology – custom robotics systems that allow for high throughput ALE experiments to be run in an automated manner. With this technology we are able to maintain upwards of 30 independent cultures in continuous exponential phase batch propagation, with real-time tracking of each culture's growth rate and precise environmental control. Compared to what can be accomplished with manually performed ALEs, we have seen equivalent fitness increases on a time scale that is nearly ten times faster [4, 5]. The ALE machines thus serve as a potent tool with which to explore the adaptive response of microbes to a wide variety of conditions.

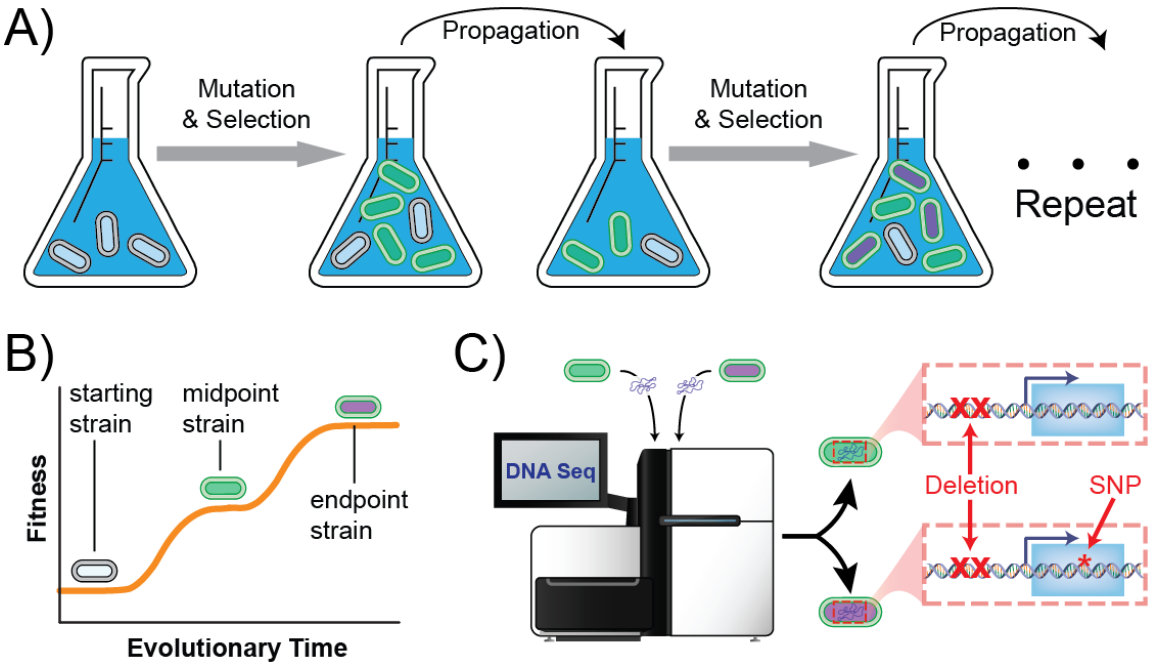



**Figure 1.1 Adaptive Laboratory Evolution** A) Microbes are cultured in a desired growth environment for an extended period of time, allowing natural selection to enrich for mutant strains (altered coloration) with improved fitness B) Evolved strains are characterized for phenotypic improvements relative to the ancestral strain, using whatever "fitness" metric is appropriate given the evolutionary environment C) Evolved strains have their DNA sequenced to reveal the adaptive mutations enabling phenotypic improvement. This example case depicts the fixation of two successive mutations targeting the same genetic region

## 1.2 Introducing the Thesis

To extensively probe the adaptive flexibility of *E. coli*, evolution to a broad range of environments is necessary. Any environment differing from the typical habitat of an enterobacterium perforce serves as a 'perturbation' towards which natural selection will begin to enrich for novel adaptive mutants. This dissertation revolves around using ALE machine technology to explore the adaptive response of *E. coli* to distinct categories of perturbation – alterations to the cell's physical environment, its chemical environment, and to its very own DNA.

A purely physical perturbation, in contrast to one of a chemical nature, inherently precludes the alteration of the chemical structure or composition of molecules in the cells' environment, or indeed the cell itself. Although at macroscopic scales there is a clear distinction between the physical and chemical environment of an organism, this division becomes blurrier at the scale of microbes – for example, although pressure is a physical attribute, to change the osmotic pressure experienced by a cell would require changing salt concentration in the culture medium, which could have unanticipated effects of a chemical nature. Perhaps the "purest" physical change a microbe naturally experiences is that of altered temperature. Temperature, i.e. the average kinetic energy of all the molecules composing and surrounding an organism, unsurprisingly has a drastic effect on survivability. Most organisms have an "optimal" growth temperature that falls within a permissive temperature range, outside of which cell death occurs. For *E. coli*, given its status as an enterobacterium, this optimal temperature coincides with human body temperature: 37 °C. Chapter 2 presents an investigation of *E. coli* adaptation to the stressful elevated temperature of 42 °C.

For laboratory *E. coli*, the "typical" chemical environment could probably be considered to be LB media, given its widespread use. However, LB is not a well-defined medium – it provides the necessary vitamins, minerals, and nutrients for robust bacterial growth, but its actual chemical composition in terms of concentrations is not quantified or consistent between batches. For this reason minimal media types are often preferred, which are produced so as to have precisely known concentrations of a minimal set of growth-permitting nutrients. There are various minimal media recipes (M9, MOPS, etc.), but the main factor distinguishing a minimal medium is the carbon source it contains.

Although there are many chemical elements necessary for survival and proliferation, catabolism of carbon compounds is the key mechanism through which energy is generated by organisms. The simple sugar glucose can be rapidly metabolized and typically allows for the fastest microbial growth, thus it is viewed as the default carbon source. Given all of this, a change of *E. coli*'s carbon source away from glucose was chosen as an ideal chemical perturbation with which to probe the adaptive evolution process. Chapter 3 presents an investigation of *E. coli* adaptation to growth on uniformly labeled Carbon-13 glucose. Chapter 4 presents an investigation of *E. coli* adaptation to growth in an environment rapidly alternating between the carbon sources glucose, xylose, glycerol, or acetate.

In contrast to physical and chemical perturbations, a significant alteration to the very DNA of an organism is not a selective pressure that existed, at least to the current extent, before humans began genetic engineering. Although changes to DNA are the reason evolution happens, the large scale changes enabled by genetic manipulation (e.g. gene knock-outs and knock-ins) are not something organisms are generally used to, and thus cell viability and robustness can be heavily compromised after such engineering, though evolving to better tolerate the newfound biological state can ameliorate things [6]. Such biological perturbations have become extensive and both commercially and academically valuable thanks to current genetic engineering techniques. A key question one might ask of genetically engineered cells is how they differentiate "self" from "non-self" when it comes to their heterologous genes. Chapter 5 presents an investigation of *E. coli* adaptation following genetic engineering to replace two important metabolic genes with foreign copies, from donor species spanning all domains of life.

Chapter 1 borrows in part from a manuscript in preparation for submission: Sandberg TE, Weng LL, Salazar MJ, Palsson BO, Feist AM. “Adaptive Laboratory Evolution: A Valuable Technique for Discovery and Industrial Biotechnology” The dissertation author is one of the authors of the review.

# Chapter 2

# Evolution of *Escherichia coli* to 42 °C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations

## 2.1 Introduction

Adaptive laboratory evolution, or ALE, has developed over the years into a potent tool for biological discovery and engineering [2]. By exploiting the inherent competition at play between organisms and the natural accumulation of mutations within a microbial population, desired phenotypic traits can be selected for without requiring *a priori* knowledge on how the traits might arise. These adaptively evolved organisms can then be subjected to whole genome resequencing, uncovering the genetic changes that enabled their phenotypic alteration. Additional data types, such as transcriptomics or metabolic uptake and secretion rates, serve to characterize the evolved strains and how they diverged from their ancestor, a divergence which must be enabled by their altered genotype. This analysis shines light on the functionality of particular genes [7] and the dynamics of the evolution process [8], increasing the biological knowledge base. While

serving as a method to perform basic scientific inquiry such as this, ALE can be an equally useful tool for applied research, pairing with synthetic and systems biology to aid in the engineering of strains [9].

ALE experiments often examine adaptation following a perturbation, either metabolic (e.g., growth on alternate carbon sources [10] or following knockout of metabolic genes [11]) or stressful (e.g., exposure to osmotic stress [12] or high ethanol concentrations [13]). However, the selective pressure guiding the adaptation can also be influenced in large part by the environment in which the strain is evolved. Evolutionary environments typically involve either batch culturing, wherein populations (often several in parallel) are serially propagated to new flasks with fresh growth medium at regular intervals, or chemostats, in which growth in a bioreactor allows for tight control of nutrient levels and other factors such as pH and oxygenation. In either case, 'fitness' ostensibly refers to a growth advantage, but this becomes more complicated by the existence of spatial or temporal inhomogeneities in the culturing environment that can lead to ecological niches. For example, in a chemostat bacteria can persist by adhering to the walls of the bioreactor [14], while batch cultures that reach stationary phase before passage can spawn subpopulations optimized for different phases of growth [15]. If the target of investigation is the method by which a cell will evolve to a particular perturbation, it can be desirable to confine 'fitness' to a single aspect, reducing the potentially confounding variables towards which a population might additionally be evolving. To this effect, batch culture serial propagation in mid-exponential phase ensures that selection occurs primarily for growth rate.

In this study we sought to examine the ALE process by adaptively evolving the wild-type mesophilic bacterium *Escherichia coli* K-12 MG1655, arguably the most highly studied and well-characterized microorganism, to constant exponential growth at a stressful elevated temperature in glucose minimal media. Ten parallel populations were evolved at 42°C using an automated system allowing for passage of batch cultures in mid-exponential phase multiple times a day, enabling many generations of growth in a relatively short time. While genes that mutate in parallel across independently evolved populations are often taken to be the likely causes for the fitness increase [16], true causal determination would require knocking-in each mutation to the starting strain in all possible combinations and comparing the resultant fitness, which would capture the individual effects of each mutation as well as their epistatic interactions. However, this quickly becomes prohibitively time consuming in strains with > 3 mutations, thus a different tool with which to probe mutational causality would be advantageous. For this reason we examined multiplex automated genome engineering (MAGE) [17] as a technique to supplement ALE experiments. After identifying the mutations occurring in the endpoints of this evolution study, MAGE was used to introduce these ALE-acquired alleles in a random fashion into the starting strain, allowing the combinatorial knock-in method to be somewhat mimicked. By competing the heterogeneous populations of genetically engineered strains against one another and determining the mutants that frequently emerged victorious, causal mutations were identified and compared with those inferred by mutational parallelism across the ALE populations.

Elevated temperature was selected as the perturbation of interest for several reasons. Firstly, to aid in the analysis of mutational causality it would be beneficial to

have more than simply one or two frequent gene targets, and adaptation to a global stress provided a diverse set of genetic changes. Additionally, a previous study [18] investigated evolution of a large number of replicates in a very similar environment, differing only moderately in ancestral strain and method by which the batch cultures were serially propagated. Comparison with this work allowed examination of the extent to which mutational parallelism persists across studies. Furthermore, examining temperature stress was desirable because two other studies, both involving metabolic perturbations [6, 19], have provided strong evidence that a large feature of evolutionary adaptation involves acclimatizing changes in gene expression back towards pre-perturbed levels. By evaluating the transcriptome of our temperature-evolved strains we sought to determine whether this trend extended to stress perturbations, which would indicate that it may be a general feature of evolution, irrespective of the nature of the perturbation.

## 2.2 Results

### 2.2.1 Evolution Process and the Endpoint Phenotypes

Ten independent populations, started from wild-type *E. coli* K-12 MG1655, were adaptively evolved in M9 minimal media supplemented with 4 grams/liter glucose at 42°C for approximately 45 days. Cultures were serially passed (~5 times per day) to flasks with fresh media once reaching a target optical density such that stationary phase was never reached and glucose concentrations were always in great excess (never dropping below 3 g/L). As mutations accrued and gained dominance within the separate populations, their fitness increased markedly relative to the ancestral strain. The populations followed different trajectories along the fitness landscape, arriving at final

growth rates on average 1.45 (± 0.06, standard deviation) fold higher than the starting point (fig. 2.1; raw fitness data shown in supplementary fig. S1). The populations underwent approximately 1500 generations of growth (supplementary fig. S2A), but because mutations occur predominantly due to DNA polymerase errors in genomic replication during cell division [20], the cumulative number of cell divisions (CCD) serves as a more meaningful scale for the time coordinate of an ALE than do generations [21]. This metric accounts for the population subsampling inherent to serial passage of cultures (supplementary fig. S2B). The CCD reached by the independent populations at the conclusion of the experiment ranged from 4.5 – 7.1 x $10^{12}$ with an average of 5.5 x $10^{12}$. With the exception of the outlier experiment #1, CCD and final population fitness were significantly correlated (Pearson's $r$ = 0.93 or 0.58 excluding or including experiment #1, respectively). This correlation is lost when using generations as the time coordinate (Pearson's $r$ = 0.35 or 0.06 excluding or including experiment #1, respectively).

Clones were isolated from each of the evolved population endpoints and subjected to further analysis to determine the physiological differences at 42°C between the wild-type and the evolved strains (table 2.1, fig. 2.2). Interestingly, the fitness (i.e., growth rate) increase of the isolated clones relative to the ancestor is noticeably lower than that of the populations (paired $t$-test, $p < 10^{-7}$). Although the discrepancy could be due to population-level altruistic interactions [22], it may result simply from the evolved populations being fully physiologically adapted to constant exponential growth after ~1500 generations under static experimental conditions. This adapted intracellular state is in contrast with the evolved clones, for which growth curves were started up from

stationary phase overnight cultures, potentially resulting in suboptimal performance due to insufficient time for re-acclimation of their protein expression machinery to the exponential growth, an effect previously documented [23]. Indeed, additional growth rate tests on the fastest and slowest growing endpoints clones (strains 4 and 9, respectively) and the populations from which they were isolated support this hypothesis, and reveal that the clone and population growth rates are in fact in excellent agreement (supplementary fig. S3). Thus, the selected clones were assumed to be representative of the dominant phenotype and genotype within the population endpoints.

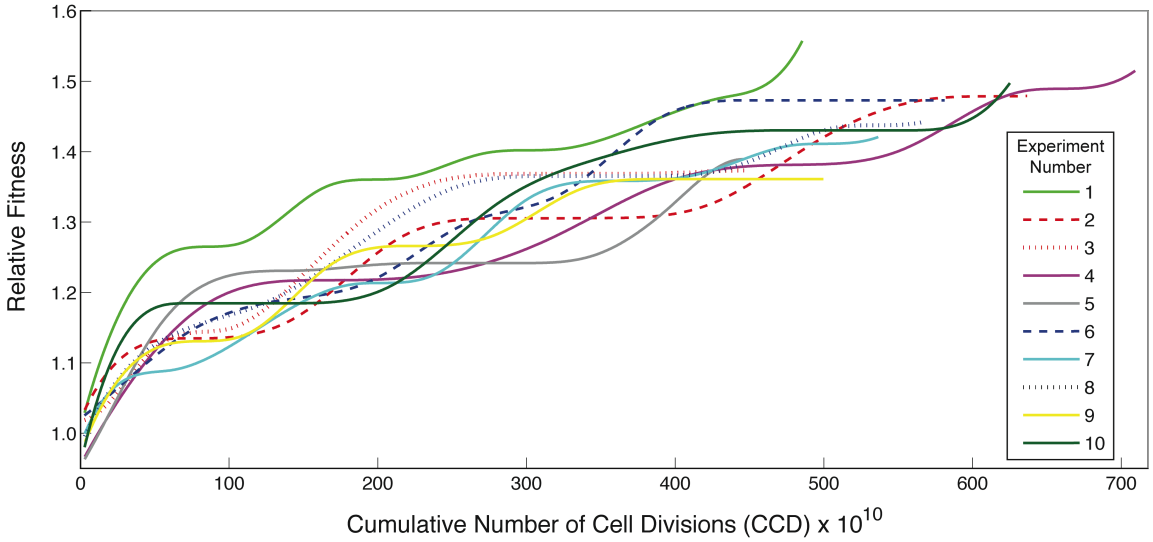



**Figure 2.1. Fitness trajectories of the evolving populations.** Plotted is the population increase in fitness relative to the initial average growth rate of the common starting strain. At the approximate CCD of 200 x $10^{10}$, the even-numbered experiments' serial passage volumes were increased 10 fold to examine the impact of CCD on overall fitness. The variability in CCD across experiments is due to this as well as fluctuations encountered in passage cell density over the course of the ALE. Dashed and dotted lines represent populations that became dominated by hypermutators, with identical colors indicating a similar hypermutator genotype.

**Table 2.1: Physiological characterization of colonies isolated from evolved population endpoints**

| Strain # | New Mutations Relative to WT | Growth Rate ($hr^{-1}$) | Relative Fitness Increase | Glucose Uptake Rate (mmol $gDW^{-1}$ $hr^{-1}$) | Acetate Production Rate (mmol $gDW^{-1}$ $hr^{-1}$) | Biomass Yield (gDW $gGlc^{-1}$) |
|---|---|---|---|---|---|---|
| WT | 0 | 0.82±0.01 | 1 | 10.2±0.2 | 5.8±0.4 | 0.45±0.02 |
| 1 | 6 | 0.95±0.03 | 1.17±0.04 | 11.8±0.9 | 5.8±1.9 | 0.45±0.04 |
| 2[a] | 34 | 0.97±0.03 | 1.19±0.04 | 15.5±1.2 | 11.0±1.6 | 0.35±0.03 |
| 3[a] | 30 | 0.92±0.04 | 1.12±0.06 | 13.7±1.8 | 12.9±0.6 | 0.37±0.06 |
| 4 | 8 | 1.03±0.01 | 1.26±0.01 | 14.3±0.3 | 10.9±0.6 | 0.40±0.01 |
| 5 | 8 | 0.94±0.05 | 1.15±0.06 | 13.9±0.7 | 10.1±0.4 | 0.38±0.04 |
| 6[b] | 41 | 0.97±0.01 | 1.19±0.02 | 15.5±1.9 | 10.9±2.5 | 0.35±0.04 |
| 7 | 8 | 0.99±0.01 | 1.21±0.03 | 14.8±2.4 | 8.0±0.6 | 0.37±0.06 |
| 8[b] | 55 | 0.95±0.03 | 1.17±0.02 | 16.0±0.7 | 13.8±1.2 | 0.33±0.03 |
| 9 | 6 | 0.92±0.01 | 1.12±0.02 | 14.3±1.7 | 11.4±0.3 | 0.36±0.04 |
| 10 | 8 | 0.98±0.02 | 1.19±0.03 | 13.6±0.7 | 9.5±0.5 | 0.40±0.03 |

[a] and [b] are hypermutator strains of the same lineage. For growth rate and relative fitness the standard deviation based on triplicate experiments is given; for other values the 95% confidence interval is given

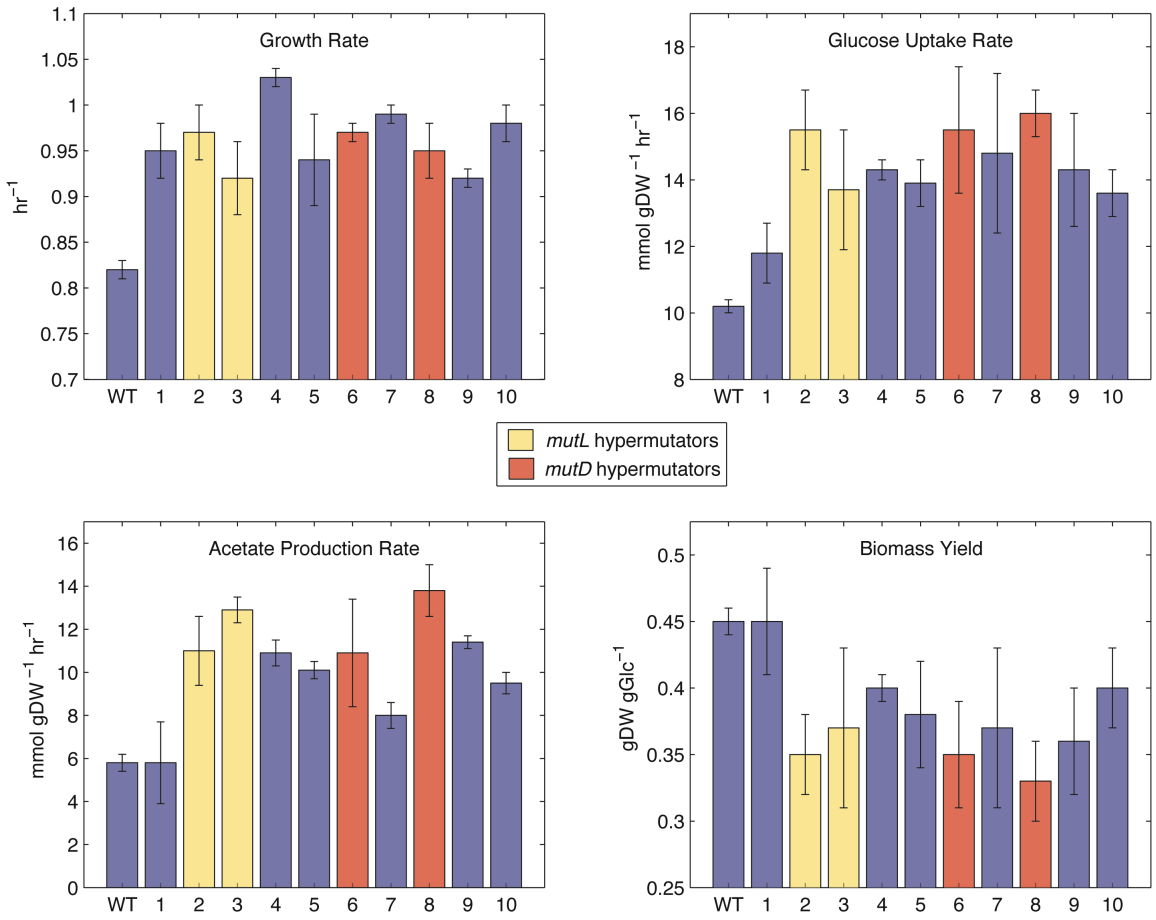



**Figure 2.2. Physiology of the temperature-evolved strains.** Strain numbers are listed below their respective bars. The error bars for growth rate represent standard deviation of three biological replicates, while error bars for the other traits represent 95% confidence intervals.

The evolved strains all displayed similar physiological changes. However, genome examination revealed that strains 2 & 3 and 6 & 8 likely shared a lineage at some point (see Mutational Analysis), thus strains 3 and 8 were omitted from the following statistics to ensure that only fully independently evolved phenotypes were considered (changing which two strains to omit did not significantly alter any values). On average the independently evolved strains increased their growth rate (μ) by 0.15 $hr^{-1}$ (equivalent to a decrease of 7.7 minutes in doubling time), increased their glucose uptake rate (GUR) by 4.0 mmol $gDW^{-1}$ $hr^{-1}$, increased their acetate production rate (APR) by 3.9 mmol

$gDW^{-1}$ $hr^{-1}$, and decreased in biomass yield by 0.07 gDW $gGlc^{-1}$ ($Y_{X/S,ss}$; calculated at steady state by dividing growth rate by GUR). Correlation plots between the pairwise combinations of these characteristics highlight the physiological divergence of the evolved strains from the wild-type, as well as the relation or lack thereof between specifics traits (fig. 2.3). There is a strong negative correlation (Pearson's $r$ = -0.94) between biomass yield and GUR among the evolved strains and a notable positive correlation ($r$ = 0.74) between APR and GUR, both of which contribute to the negative correlation between biomass yield and APR ($r$ = -0.76). These results imply that the strains adopted a similar phenotypic change, but to a varying extent, of increased glucose uptake at the cost of increased acetate overflow metabolism [24], utilizing this greater metabolic flux not to create biomass but rather to generate more energy. Notably, growth rate itself was not correlated with any of the other traits (no Pearson's $r$ with a magnitude above 0.22). These physiological trends replicate what is observed upon evolution to glucose minimal media at 37°C [5].

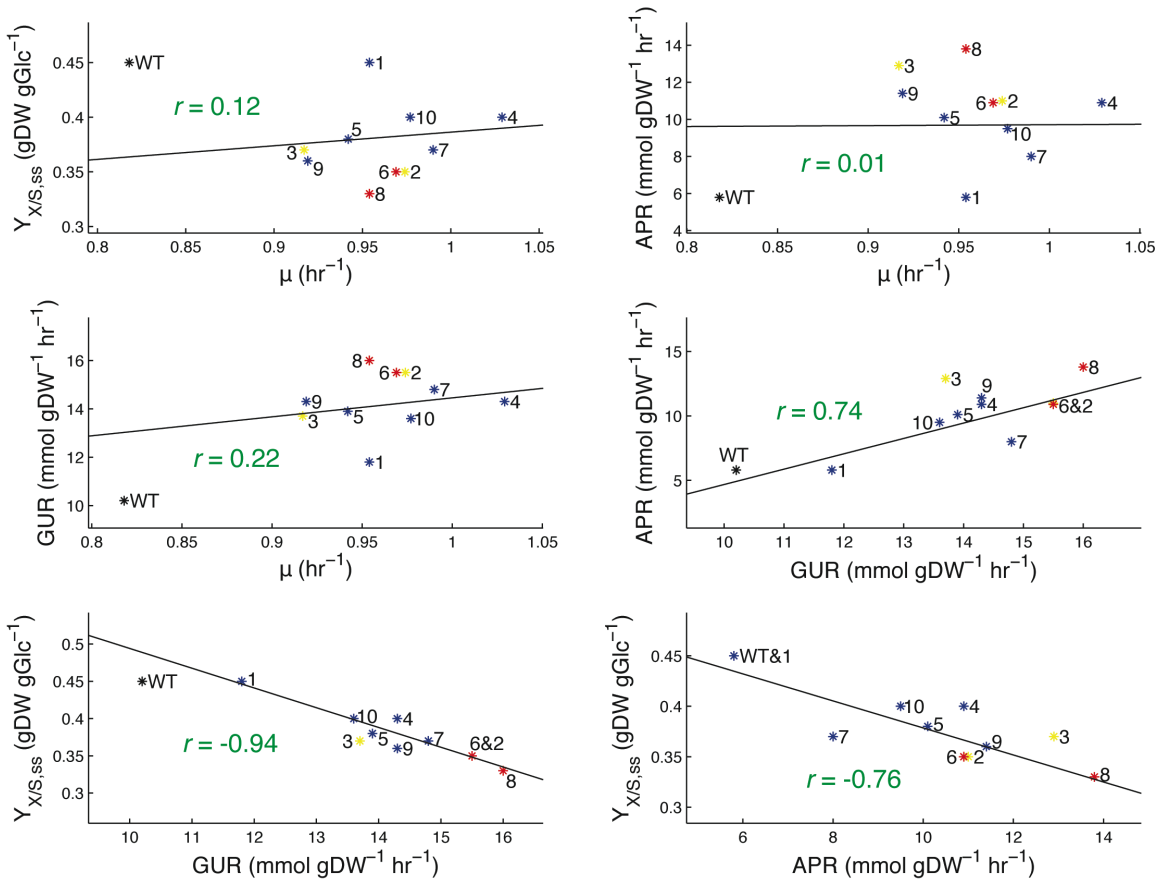



**Figure 2.3. Pairwise correlation plots between the physiological traits of the evolved strains.** Strain numbers are listed next to their respective points, with the *mutL* hypermutators in yellow and the *mutD* hypermutators in red. The least squares linear best-fit line to the evolved strains (excluding the wild-type starting strain from the fit as well as strains 3 and 8, so that only fully independently evolved phenotypes are considered) is overlaid along with its Pearson correlation coefficient. ($Y_{X/S,ss}$ = biomass yield at steady-state, μ = growth rate, GUR = glucose uptake rate, APR = acetate production rate)

## 2.2.2 Mutational Analysis

Whole genome resequencing was performed on isolated clones from each experiment in order to investigate the genetic basis underlying their phenotypic changes. A total of 161 unique *de novo* mutations relative to the starting strain were found across all 10 endpoints (supplementary dataset S1), with a number of these being shared among two or more of the strains, or occurring within the same genes.

**Table 2.2: Recurring mutations identified across the ALE endpoint strains**

| Gene | Specific Function | Class[a] | Mutation | Protein Change | Strain Number(s) |
|---|---|---|---|---|---|
| *pyrE/rph* | Orotate phosphoribosyl-transferase/RNase PH | M | 82 bp deletion | frameshift | 1, 2, 3, 4, 5, 7, 9, 10 |
| *rpoC* | RNA polymerase subunit | R | SNP | A734V (GCG→GTG) | 1, 2 |
| | | | SNP | Q1367* (CAG→TAG) | 5, 7, 10 |
| *pykF* | Pyruvate kinase I | M | SNP | ydhZ/pykF intergenic (-498/-59) | 3 |
| | | | SNP | T278S (ACC→TCC) | 10 |
| | | | SNP | K286* (AAA→TAA) | 4, 5, 7 |
| *rne* | Ribonuclease E | R | SNP | D415N (GAC→AAC) | 6/8 |
| | | | SNP | H243Y (CAT→TAT) | 3 |
| | | | SNP | G124S (GGT→AGT) | 9 |
| | | | SNP | V19A (GTA→GCA) | 2 |
| *ygaH/mprA* | L-valine efflux transporter/MprA repressor | R | SNP | intergenic (+77/-14) | 6/8 |
| | | | SNP | intergenic (+80/-11) | 10 |
| | | | SNP | intergenic (+81/-10) | 4, 5 |
| *mlaE* | Phospholipid ABC transporter | C | SNP | L107F (TTG→TTT) | 4, 5, 7 |
| *yfdI* | Predicted inner membrane protein | C | 1 bp insertion | frameshift (519/1332 nt) | 8 |
| | | | SNP | Q186* (CAA→TAA) | 6 |
| | | | 1 bp deletion | frameshift (1274/1332 nt) | 3 |
| *nagC* | PTS regulator | C | SNP | C264S (TGC→AGC) | 6/8 |
| | | | 1 bp deletion | frameshift (218/1221 nt) | 2 |
| *hns/tdk* | H-NS regulator/ thymidine kinase | R | SNP | intergenic (-22/-583) | 6/8 |
| | | | Insertion sequence | intergenic (-111/-486) | 9 |
| *hfq* | RNA binding protein | R | SNP | D9A (GAT→GCT) | 1, 7 |
| *nagA* | N-acetylglucosamine-6-phosphate deacetylase | C | SNP | G265D (GGC→GAC) | 2 |
| | | | 21 bp deletion | In-frame (381-401/1149 nt) | 1 |
| *secD* | Membrane protein channel component | C | SNP | R181L (CGC→CTC) | 8 |
| | | | SNP | G499G (GGC→GGT) | 2 |
| *dinQ/arsR* | Toxic membrane peptide/ metal-responsive regulator | C | SNP | intergenic (-209/-486) | 8 |
| | | | 1 bp deletion | intergenic (-305/-390) | 3 |
| *ilvL/ilvX* | Leader peptide regulating isoleucine & valine biosynthesis operons | M | Double SNP | intergenic (+47/-39) | 6 |
| | | | SNP | intergenic (+48/-39) | 8 |

[a] M, R, C = Metabolic, Regulatory, Cell envelope

The emergence of several 'hypermutators,' a recurring feature of many ALE experiments [25], accounted for the majority of these unique mutations. Four of the ten strains were hypermutators, possessing on average 40 mutations, while the remaining six 'non-mutators' had an average of 7 mutations. The resequencing results indicate that two of the four hypermutators likely resulted from unanticipated cross-mixing between the evolving populations, thus only two hypermutators can be said to have occurred independently: one likely due to a SNP in *mutL* (strains 2 and 3) [26] and the other due to a SNP in *mutD*, also known as *dnaQ* (strains 6 and 8) [27]. Analysis of the non-mutator genotypes does not indicate that they suffered from similar occurrences of cross-mixing (see Materials and Methods).

The observed mutations in the clones isolated from each population were compared. Fourteen genes or intergenic regions were found to mutate in parallel across two or more of the evolved strains (discounting those mutations shared due to cross-mixing), which fell into three general functional categories: mutations affecting metabolism, regulation, or the cell envelope (table 2.2). Intergenic mutations were categorized based on their position relative to the genes (e.g., SNPs downstream of one gene and upstream of another likely change expression of the latter) and transcriptomic data obtained from RNA-seq analysis. The key mutations within each of the three functional categories are now described:

1) Metabolic Mutations: Only three metabolic genes were found to mutate in more than one strain, but two of these occurred in half or more of the strains. Foremost among all mutations, regardless of category, is an 82 bp deletion between *pyrE* and *rph* that occurred in every strain except the two *mutD* hypermutators. This mutation does not

appear to be specifically temperature-related, having been observed in several other ALE studies on adaptation to lactate- or glucose-supplemented minimal media at 30°C or 37°C, respectively [28, 29], and is hypothesized to relieve a defect in pyrimidine biosynthesis present in the starting strain [30]. The second most predominant metabolic mutation was in *pykF*, or pyruvate kinase I, which experienced one intergenic SNP (accompanied by a 2.7 fold down-regulation in gene expression) and two different nonsynonymous mutations that may cause PykF inactivation through premature truncation of the enzyme (K286*) or alteration of a putative substrate binding residue (T278S) [31]. As with *pyrE/rph* this is likely not a temperature-specific beneficial mutation, given that PykF inactivation is a recurring feature of *E. coli* adaptation to glucose minimal media, hypothesized to allow for increased glucose uptake by decreasing the metabolism of phosphoenolpyruvate to pyruvate [29, 32, 33].

2) Regulatory Mutations: Five of the fourteen recurring mutations were found in regulatory genes, with three of these occurring in half of the strains: *rpoC*, *rne*, and *ygaH*/*mprA*. RpoC is a subunit of the RNA polymerase complex, which, given its ability to function as a global regulator, is a frequent target of mutations in bacterial adaptation [34, 35]. Similar ALE-identified *rpoC* mutations have been found to increase the general metabolic efficiency of *E. coli* grown in minimal media [7, 36] or are inferred to adapt it to higher temperatures [18]. The ribonuclease *rne* had 4 different SNPs, the most diverse set of mutations observed in any one gene in this study. Rne is an essential enzyme involved in rRNA and tRNA processing and, as a

key component of the RNA degradosome, is the rate-limiting or sole degrader of many transcripts [37]. A previously discovered *rne* mutant was nonviable at 44°C and significantly defective in rRNA/tRNA processing as well as mRNA degradation, but could be rescued at the elevated temperature by SNPs in the N-terminal catalytic domain (in which all four ALE-acquired SNPs occur) that restored the rRNA/tRNA processing to wild-type levels but did not undo the 2-3 fold decrease in mRNA decay rate [38]. Relevantly, in prokaryotes the stability of mRNA is directly correlated with the optimal growth temperature of the organism [39], suggesting that adaptation strategies to increased temperature might include increasing the stability of mRNA transcripts. Taken together, this implies that the *rne* mutations found in this study may function to improve the enzyme's rRNA/tRNA processing capabilities at 42°C without likewise improving its endonuclease efficacy. It is also of note that a SNP in *hfq* was observed in two strains, and Hfq binding can prevent mRNA degradation by Rne (fig. 2.4A) [40]. The three different intergenic *ygaH*/*mprA* SNPs all occurred between 10-14 nucleotides upstream of the *mprA* start codon, likely influencing translation efficiency by modulation of the ribosomal binding site [41]. MprA is a transcriptional repressor for a number of genes that code for multidrug resistance pumps [42]. Although drug resistance is not a factor in this evolution, increased expression of the pumps leads to altered membrane flux for a variety of compounds [43], thus this regulatory mutation could also perhaps be classified as cell envelope-related.

3) Cell Envelope Mutations: Despite no single cell envelope-related gene being mutated in more than 3 of the 10 endpoints, there was nevertheless a clear selective pressure on the cell envelope in general, with 6 of the 14 recurring mutations falling into this category. These range from phospholipid transporters to membrane proteins and channels to genes involved in the levels of cell envelope components. Note that although *nagC* and *nagA* could be classified as regulatory and metabolic, respectively, their role in the recycling of cell wall peptidoglycan (fig. 2.4B) is responsible for their classification as cell envelope-related mutations [44]. This category of mutations is a feature of ALEs in general [28, 45] and in this particular study may help the envelope maintain its essential physical properties at the elevated temperatures of the experiment [46].

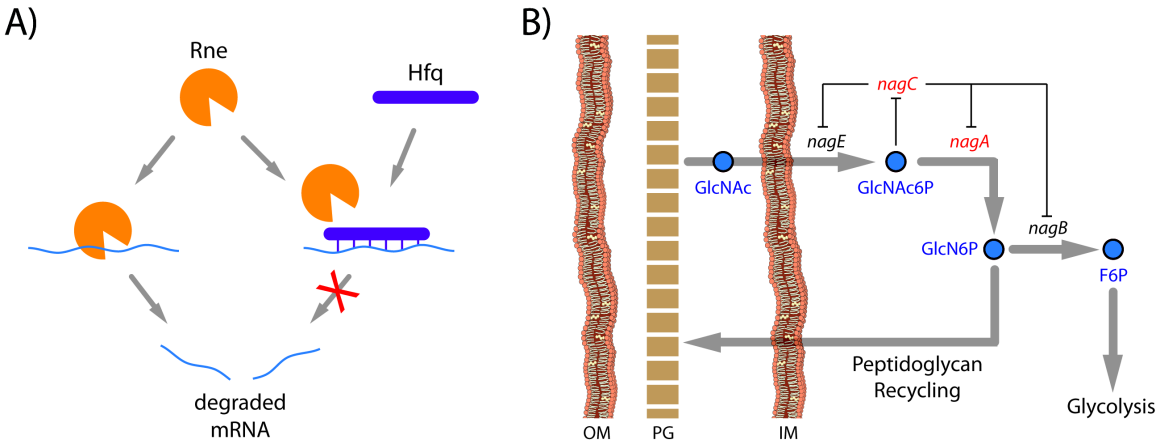



**Figure 2.4. Repeatedly mutated genes with related functionality.** (A) Both *rne* and *hfq*, mutually involved in mRNA degradation, mutated in multiple of the temperature-evolved ALE strains. (B) Genes in the *nag* operon facilitate uptake of GlcNAc from the periplasmic space, which can be channeled away into glycolysis or reincorporated into the peptidoglycan of the cell wall. Genes in red (*nagC*, *nagA*) were recurring mutational targets. (OM = outer membrane, PG = peptidoglycan, IM = inner membrane, GlcNAc = N-acetylglucosamine, GlcNAc6p = N-acetylglucosamine-6-phosphate, GlcN6P = glucosamine-6-phosphate, F6P = fructose-6-phosphate)

Although a number of ALE studies examining adaptation to increased temperature have been performed to date [18, 32, 47, 48], the one by Tenaillon *et al.* serves as an ideal point of mutational comparison given the significant similarity in culture environment between our two studies and the large number of independent lines they evolved, providing a statistically significant basis against which to compare the mutations observed herein. However, despite the similarity in evolution environment, the difference in mutational frequency is extremely pronounced. Of the fourteen genes in this study that mutated in two or more strains, only four were found to mutate in *any* of the 114 evolved lines sequenced by Tenaillon. Both *rne* (4 different SNPs across 5 strains) and *ygaH*/*mprA* (3 different SNPs across 5 strains) had a mutation in only 1 out of the 114 lines. Mutations in *rpoC* and those in or around *ilvL* were the only noticeably recurring feature: 5 out of 10 and 2 out of 10 *rpoC* and *ilvL* mutants, respectively, in the evolved strains of this study, vs. 21 out of 114 and 29 out of 114 identified in the Tenaillon strains. Even given the relatively small sample size in this study, to have greater than 20% mutational infiltration of 12 genes in one instance and less than 1% in another implies a substantial difference in evolutionary trajectory. Though the recurring mutations in this study (table 2.2) overlap poorly with those identified by Tenaillon, there is slightly more overlap between the functional units determined to be significant in their study (possessing 5 or more mutations across the 114 strains) and the mutations observed in this work: 9 of the 26 functional units share one or more mutated genes, despite most mutational overlap occurring in only a single one of our evolved strains (supplementary table S1).

### 2.2.3 Analysis of Mutational Causality Using MAGE

Multiplex automated genome engineering (MAGE) [17] was utilized to examine causality of the various mutations identified in the evolved strains. To limit the scale of the experiments, only those mutations found in the 6 non-mutator strains were selected as targets to introduce via recombineering. This yielded 31 distinct mutations between the strains, resulting in 29 unique oligos with which to perform recombineering (two of the mutations were intergenic insertion sequences, infeasible for use in MAGE due to their size, while insertion sequences occurring within genes were assumed to function equivalently to knock-outs). These oligos were organized into seven distinct pools – six pools which contained only those oligos corresponding to the mutations found within each of the non-mutator evolved strains (strains 1, 4, 5, 7, 9, and 10), and one pool which contained all 29 oligos. Nine rounds of recombineering were performed on the starting strain for each of these 7 pools, after which serial growth and passage of the cultures occurred to enrich for those strains existing within the highly heterogeneous starting populations that were most fit for growth at 42°C. Multiple colonies were isolated from each of these enriched populations and subjected to whole genome resequencing (supplementary dataset S2).

The most frequently observed MAGE mutations (occurring in at least three of the enriched strains, with a frequency > 25%) are given in table 2.3. Frequency for a gene is defined as the number of resequenced strains which possessed a mutation divided by the number of resequenced strains which potentially could have possessed it given the pool of oligos used in recombineering. This recapitulates what was observed in the evolved ALE strains – there is excellent agreement between the most frequent MAGE and ALE

mutations, keeping in mind that the use of only non-mutator strain mutations in MAGE explains the absence of *yfdI*, *nagC*, *secD*, *dinQ/arsR*, *ilvL/ilvX*, and *hns/tdk* (fig. 2.5). This leaves *proQ*, a regulator of the membrane transporter ProP, as the sole disagreement between tables 2 and 3, which mutated in only ALE strain 7 (a missense SNP, Q216P) and yet was found in 9 of 19 relevant MAGE strains. These results reinforce the conclusions drawn from examination of mutational parallelism across the independently evolved ALE strains – that the identified genes are likely key targets for adaptation to growth at 42°C in glucose minimal media.

**Table 2.3: The most significantly recurring genes following enrichment of MAGE strains**

| Gene | MAGE frequency | ALE frequency | New off-target mutations[a] |
|---|---|---|---|
| *pyrE/rph* | 0.88 | 0.80 | 26 (including 9 frameshifts) |
| *ygaH/mprA* | 0.86 | 0.50 | 1 (intergenic 1bp deletion) |
| *rpoC* | 0.79 | 0.50 | 2 (I1357I, N762S) |
| *pykF* | 0.55 | 0.50 | 0 |
| *hfq* | 0.50 | 0.20 | 0 |
| *proQ* | 0.47 | 0.10 | 1 (G212G) |
| *nagA* | 0.36 | 0.20 | 2 (G127G, H129N) |
| *rne* | 0.33 | 0.50 | 0 |
| *mlaE* | 0.27 | 0.30 | 1 (L99L) |

[a] Protein alterations given in parentheses

On average, enriched strains possessed 4 mutated genes targeted by recombineering and 3 mutations in secondary locations throughout the genome. However, several of the targeted genes experienced mutations that differed from what would be expected given the design of the recombineering oligos. With the exception of one, all of these novel, 'off-target' mutations occurred within the 70 bp region introduced into the genome by allelic replacement and thus likely resulted from incorporation of mis-synthesized oligos containing erroneous bases, as has been observed with MAGE

previously [17]. Nevertheless, examining these unintentional mutations that were able to fix in the post-enrichment populations provides insight into the nature of the causal genetic changes.

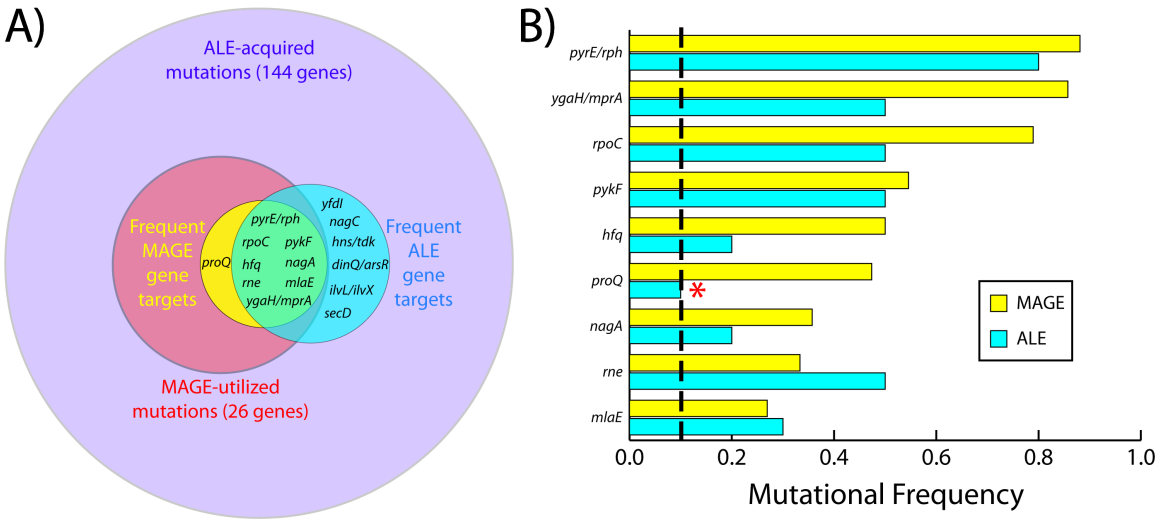



**Figure 2.5: Mutational agreement between the MAGE strain enrichment and ALE experiment results.** (A) A subset of the novel alleles acquired via evolution were randomly introduced into the ancestral strain with MAGE, and these MAGE strains were competed against one another in the enrichment process. 'Frequent ALE gene targets' refer to genes that mutated in at least two of the ten ALE endpoint strains, while 'Frequent MAGE gene targets' refer to those that mutated in at least 25% of the relevant post-enrichment MAGE strains. Circle areas are proportional to the number of genes contained within each category. (B) Frequency comparison between MAGE and ALE strains for the 9 most frequent MAGE gene targets. The only disagreement occurs for *proQ* (marked by *), which was not a frequent ALE gene target. The vertical dashed line corresponds with 10%, the frequency for a gene mutating in only a single one of the ten ALE strains.

In the case of *pyrE/rph*, 26 new SNPs and indels were found, and in no instance did a strain possess some combination of *rph* or *pyrE/rph* mutations that did not include the introduction of a shifted *rph* stop codon. Rph is naturally defective in the *E. coli* strain used in this study [30], so alterations to its coding sequence should be phenotypically neutral. These mutations support the mechanism put forth previously [28], whereby a frameshift that moves the *rph* stop codon closer to the *pyrE* attenuator loop allows for

improved regulation of *pyrE* expression. If these *rph* frameshifts and nonsense SNPs all yield roughly equivalent outcomes, then this begs the question of why only the same 82 bp deletion was observed in the ALE strains. This disparity in variation can be explained by the different acquisition method of ALE vs. MAGE mutations. The 82 bp *pyrE/rph* deletion is flanked by two 10 bp repeats, thus its prevalence in the adaptively evolved strains may be due to the relative frequency of DNA polymerase slippage during DNA replication [28, 49], and once this fixes within the populations there is no longer a selective force for continued genetic alteration in the region. This mutational bias is decreased under MAGE conditions – even with only a small fraction of oligos with synthesis errors, a variety of mutations can be acquired across each round of recombineering before the populations are subjected to growth-based selection in the form of enrichment, negating the ease-of-acquisition benefit the 82 bp deletion has in gaining dominance.

Unlike with *pyrE/rph*, where most frameshifts will be beneficial and gene inactivation is not a concern, one would expect that in the other MAGE gene targets the introduction of mis-synthesized oligos would in most instances result in a fitness decrease by altering the amino acid sequence of a functioning protein. This fitness decrease would render the mutants unable to survive the enrichment process, yielding few off-target mutations within the relevant genes of the enriched MAGE strains. The data support this conclusion; other than the 26 new *pyrE/rph* mutations, only 8 off-target mutations were found across all other MAGE-targeted genes in the resequenced strains, with more than half of these resulting in no protein sequence alteration (one intergenic 1 bp deletion, four synonymous SNPs, and three missense SNPs). There is no reason that the oligos used for

*pyrE/rph* should be more erroneous than those for any other gene, thus the difference in number of off-target mutations that survived enrichment highlights the relative specificity of the ALE-identified mutations for effecting a fitness increase. However, *pykF* experienced no off-targets despite its two separate oligo-introduced mutations both potentially leading to gene inactivation, either by altering a potential binding residue or truncating the protein. It may be that these mutations only decreased the enzymatic activity of PykF rather than completely eliminating it, and reduced PykF functionality happens to be more beneficial than the complete inactivation that would likely result from the introduction of random, mis-synthesized oligos. It should also be noted that one off-target mutation, a SNP in *rpoC* (N762S), fell outside of an oligo-targeted region and thus may have arisen independently.

The resequenced MAGE strains were subjected to growth rate tests to ensure that they were in fact adapted to growth at 42°C. The tests revealed two strains that, despite possessing mutations and being picked from enriched populations, did not grow faster than the starting wild-type strain. These two strains were precluded from the MAGE mutational frequency analysis, but nevertheless highlighted an interesting feature: all adapted strains possessed mutations in *pyrE/rph* and/or *rpoC*, and the two un-adapted strains were the only ones to have neither.

### 2.2.4 Transcriptomic Profiling via RNA-seq

RNA-seq was performed to examine the global shifts in gene expression resulting from the altered genotypes of the evolved strains as compared to the wild-type ancestral strain at 42°C. To complement this analysis, the expression shift of the wild-type strain

when grown at 42°C vs. 37°C was also determined (supplementary dataset S3). Figure 6A shows a heat map of fold changes for the 1208 genes deemed to be significantly differentially expressed ($q$-value < 0.05 [50]) in the wild-type strain when grown at the higher temperature. The differentially expressed wild-type genes and their pattern of up- and down-regulation are in good agreement with a previous study examining differential expression in *E. coli* after growth at 43°C [51], including such features as up-regulation of heat shock proteins (e.g., ClpB, DnaK, GrpE, and GroL, among others) and sulfur metabolism genes (*cys* genes), and down-regulation of genes involved in flagellar synthesis (*fli* and *flg* genes) and putrescine catabolism (*puu* genes).

In most of these cases, and indeed as a general trend across many of the genes (fig. 2.6A), the mutations of the evolved strains served to reverse the heat-induced transcriptional shift, restoring the expression state back towards the levels of the wild-type at 37°C. On average, each evolved strain had restorative shifts for 73% of the 1208 genes, and principal component analysis provides an additional means of visualizing these evolved shifts in gene expression (supplementary fig. S4). Such restoration has been documented in two other instances; in one study examining evolution of *E. coli* onto lactate or glycerol as the sole carbon source as opposed to glucose [19], and in another examining evolution of *Methylobacterium extorquens* following replacement of a native central metabolism reaction with a functionally analogous, heterologous pathway [6]. In both cases there was widespread restoration of expression back to the wild-type levels, highlighted by the less common 'reinforcement' of a change (i.e., down- or up-regulated genes at the start of evolution becoming increasingly down- or up-regulated following evolution, respectively). This reinforcement, when occurring across multiple of the

independently evolved strains, was taken as evidence for the reinforcing shift being an important factor in their increased fitness. In this study, highly parallel (occurring in at least 8 of the 10 evolved strains) transcriptional reinforcement occurs in 101 genes, compared to 703 genes that experience highly parallel restorative shifts. It should be noted that expression levels can be influenced by growth rate-dependent effects [52], but this cannot explain the observed transcriptional restoration given that the growth rate of the wild-type strain at 42°C is higher than at 37°C (0.82 $hr^{-1}$ vs. 0.7 $hr^{-1}$), both of which are lower than the evolved strains.

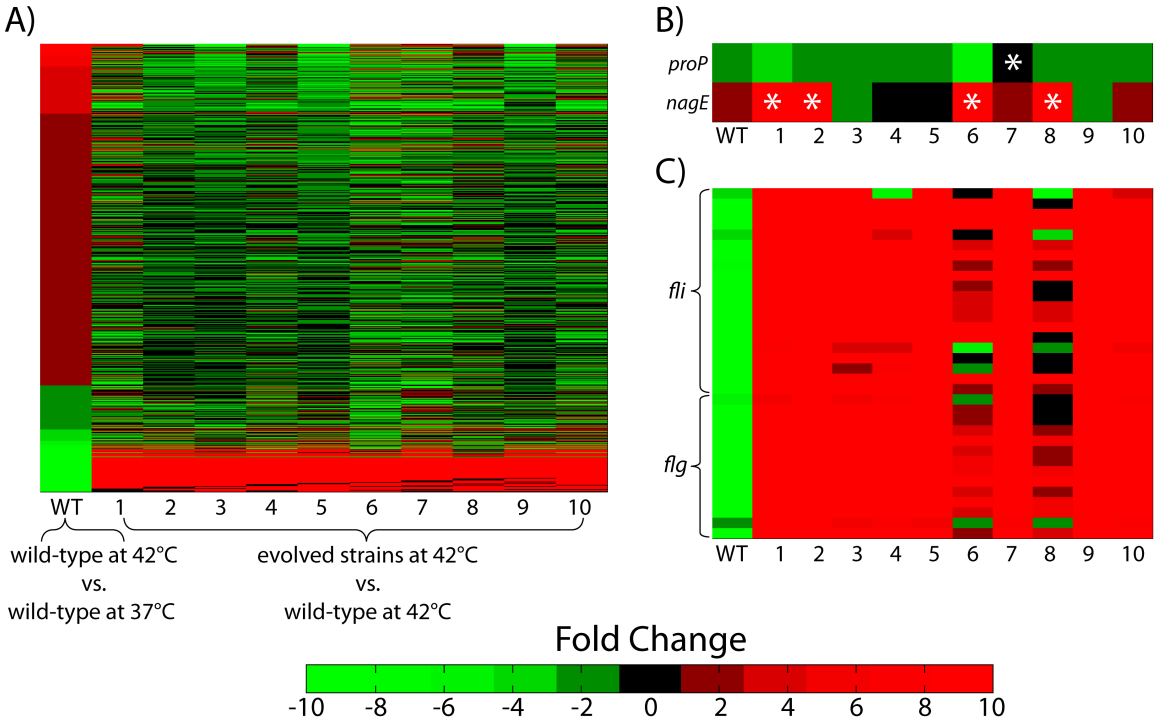



**Figure 2.6. Heat maps of differential expression between the ancestral and evolved strains.** (A) The 1208 genes which are significantly differentially expressed ($q$-value < 0.05) in the wild-type after growth at 42°C, rank-ordered. (B) Mutations acquired by the evolved strains (relevant strains marked with *) could either serve to resist (*proP*; strain 7) or aid in (*nagE*; strains 1, 2, 6, and 8) the reinforcement of the WT expression shift. (C) Widespread up-regulation of flagellar genes after evolution, partially resisted by two strains (6 and 8) possessing similar genotypes.

The reinforcing shifts in gene expression across the temperature-evolved strains were examined in more detail. Evidence was found both supporting and refuting the assumption that these reinforcements point to key mechanisms of adaptation. Supporting data include the expression shifts observed in *nagE*, which increased 1.6 fold in the wild-type at 42°C followed by a further increase between 15-71 fold in evolved strains 1, 2, 6, and 8, in stark contrast with the remaining strains which changed in expression by no more than 1.6 fold (fig. 2.6B). The strains exhibiting the strongly reinforcing increase in *nagE* expression are those possessing mutations in *nagC* and/or *nagA*, both determined to be significant gene targets in the adaptation to increased temperature (table 2). These mutations are likely responsible for the observed *nagE* shifts, given the genes' mutual involvement in peptidoglycan recycling [44]. This finding is similar to the strongly reinforced up-regulation of *pntAB* observed by Carroll and Marx in the *M. extorquens* evolved strains, which additional analysis showed to be a major contributor to their fitness increase, and the discovery of mutations within this operon.

In contrast with *nagE*, expression shifts in *proP* serve as an example of a reinforcing shift seemingly acting against the best interest of the cells. Upon initial growth at 42°C the wild-type strain decreases in *proP* expression by 1.5 fold, a decrease which is reinforced uniformly across the ten evolved strains. However, strain 7 exhibits the smallest expression decrease (an insignificant -1.2 fold, compared to an average of -2.7 fold across the other nine strains) and is also the strain that acquired a SNP in *proQ*. Though not appearing in multiple of the ALE endpoint strains, the *proQ* mutation was found to be significant in the MAGE enrichment analysis and is likely responsible for the greater comparative *proP* expression, given that ProQ's only known function is as a

regulator of ProP levels [53]. This implies that the reinforced change in expression of *proP* is actually detrimental to growth at increased temperatures, enough to allow for fixation of this mutation that combats the change and provides a fitness increase of suitable magnitude to repeatedly survive MAGE enrichment. However, although no epistatic interactions could be inferred from the MAGE results (the *proQ* mutation did not appear solely in the presence of any other single mutation), the possibility cannot be discounted that different expression shifts in *proP,* or indeed any gene, could have different effects in different strains.

In the same way that the reinforced expression shifts may not all necessarily aid in the fitness of the cells, so too is this the case for the restorative shifts. The most noticeably counterintuitive restorative shift is the massive up-regulation of flagellar genes in the evolved strains following their initial down-regulation in the wild-type upon growth at 42°C (fig. 2.6C). In the well-mixed environment of the ALE experiment there is no evident need for motility, so increased expression of these genes should incur a significant energetic cost [54] while providing no apparent benefit. Indeed, previous ALE studies, regardless of culturing temperature or carbon source in the growth medium, have yielded strains with mutations that led to the down-regulation of flagellar genes [5, 19, 55]. Interestingly, strains 6 and 8 (the two *mutD* hypermutators) appear to have somewhat mitigated this flagellar up-regulation in the same way that strain 7 did for *proP* down-regulation. The SNP 22 base pairs upstream of the *hns* start codon that strains 6 and 8 share is a potential candidate for the genetic change behind their outlying behavior, given *hns*'s part in the regulation of flagellar genes [56]. Taken together, the data suggest that

the evolved expression shifts, whether restorative or reinforcing, can sometimes be actively detrimental instead of beneficial.

Other than the influence of the *nagA/nagC* and *proQ* mutations on transcription of *nagE* and *proP*, respectively, the recurring ALE mutations (table 2) have only subtle, if at all discernible, effects on expression levels relative to the other endpoint strains lacking these mutations. This indistinct mutational influence on transcription includes both the genes that mutated, as well as separate genes that might logically experience an expression shift as a result (e.g., a *proQ* mutation causing *proP* shifts). Even the intergenic mutations, which are more likely to cause expression shifts than are protein sequence alterations [57], fail to noticeably distinguish themselves from the strains without the mutations, and can be inconsistent in their influence. For example, although the strain possessing a SNP 59 base pairs upstream of the *pykF* start codon has the greatest decrease in *pykF* expression (-2.7 fold; the second largest decrease being -2.2 fold), it fails to stand distinctly apart from the remaining strains (*t*-test, $p > 0.16$). Similarly, although 8 of the evolved strains share the identical 82 bp *pyrE/rph* deletion that should influence *pyrE* expression [28], their fold changes for the gene ranged between -3 and +5. Isolating the transcriptional effects of any individual mutation is clearly complicated by the presence of other genetic changes within the evolved strains. Elucidation of the expression shifts enabled by specific mutations would likely require the creation and transcriptomic characterization of single knock-in mutants.

## 2.3 Discussion

In this study, adaptive laboratory evolution of 10 replicate cultures was performed under carefully controlled conditions to yield *E. coli* strains selected solely for faster exponential-phase growth on glucose minimal media at 42°C. These strains were physiologically characterized, transcriptomically profiled, and subjected to whole genome resequencing to determine the genetic basis of their adaptation. Selected mutations were knocked-in to the ancestral strain in random combinations via MAGE to gain further insight into the causal genetic changes behind the acquired fitness increase. It was found that (1) frequency-based assignment of causality to mutations is largely consistent across the naturally evolved and genetically manipulated strains; (2) the path of genetic adaptation is greatly influenced both by the genotype of the starting strain and the conditions under which the evolution is performed; and (3) a variety of mutations leads to the same general trend of phenotypic and transcriptional convergence among the evolved strains, highlighted by the occasional counterintuitive shift in gene expression.

Although the number and variety of mutations capable of conferring a growth advantage in any particular environment may be nearly limitless, those which strongly contribute to an improved phenotype have a much greater probability of fixing within the evolving populations than do detrimental, neutral, or slightly beneficial mutations. Thus, genes mutated in parallel across independently evolved cultures stand out as likely causes for the observed fitness increase, rather than simply being 'hitchhiker' mutations [16]. This frequency-based analysis pointed to 14 different genes, fewer than 10% of the 144 genes mutated across all strains, as the foremost genetic targets for adaptation to growth on glucose at 42°C. A subset of these 144 altered genes, encompassing the mutations acquired by the six strains with few (≤ 8) genetic changes relative to the common

ancestor, were genetically engineered into the starting strain in random combinations with the techniques of MAGE. Repeated culturing of the highly heterogeneous MAGE populations enriched for the most fit strains existing within them, and genome resequencing revealed the genetic changes possessed by these strains. It was found that frequency-based analysis of the mutated genes within the MAGE strains perfectly recapitulated the likewise-determined key genes of the ALE strains (at least, for those genes which fell within the subset used in the genetic engineering) and pointed to only one additional key gene target that did not happen to mutate in parallel in the ALE. Additionally, the unintentional inclusion of oligos with synthesis errors in the MAGE allelic replacement [17] highlighted the effects the evolved mutations had on the genes; many unintentional off-target mutations accumulated if fitness was positively or neutrally affected by inactivation of the gene in question, but off-targets were generally lacking or synonymous if the mutated gene likely retained some measure of functionality. While not apparent in this study, MAGE analysis would also serve to highlight the existence of positive epistasis between acquired mutations if they predominantly showed up together in the post-enrichment strains, implying that their combined fitness benefit was greater than either individually. MAGE thus presents itself as a useful tool for reaffirmation of mutational causality following an ALE experiment.

Though the excellent agreement in mutational frequency between ALE and MAGE strains implies that these recurring mutations would similarly appear in other *E. coli* temperature evolution experiments, comparison with the work of Tenaillon *et al.* revealed that for the most part this is untrue. Of the 14 recurring ALE mutations identified herein, ten experienced zero mutations across all 114 lines evolved by

Tenaillon, while two mutated in only 1 of the 114 lines. To understand where this discrepancy arises, it is important to classify the ways in which the two studies differ. Firstly, in this study K-12 MG1655 is used as the ancestral strain, differing from the B REL1206 strain used by Tenaillon (a descendant of B REL606 evolved for 2,000 generations on glucose minimal media at 37°C [58]). This strain difference provides a clear explanation for several of the mutational discrepancies: the 82 bp *pyrE/rph* deletion widespread among the strains generated here is specific to MG1655, thought to relieve an inherent defect in pyrimidine biosynthesis [28], and Tenaillon does not observe *pykF* mutations because B REL1206 already contains an inactivating mutation in *pykF*, precluding it from further selective forces. The use of different starting strains also provides a possible explanation for the lack of *rne* mutants found by Tenaillon. Rph is naturally defective in MG1655 and is known to interact *in vivo* with Rne [59], so it may be that certain detrimental effects resulting from elevated temperature do not manifest themselves when Rph and Rne are able to form functional assemblies, but the absence of Rph places pressure upon Rne to develop compensatory mutations for improvement of its functionality. On top of these mutational discrepancies that can be intuited, there are doubtless those for which explanations are not obvious; for example, several beneficial *rpoB* mutations identified in the Tenaillon study were actually detrimental to fitness, or potentially even lethal, when introduced into MG1655 [60].

In addition to the different starting strains, differences in experimental methodology are likely to have a significant effect. Although evolved at essentially identical temperatures (42°C vs. 42.2°C) and on the same carbon source, the environments under which evolution occurred were nevertheless quite different. Whereas

in this study constant exponential phase growth was maintained and all nutrients were always in great excess, thus allowing for selection based solely on growth rate, the methodology of Tenaillon mimics that of the long-term evolution experiment (LTEE) [4] – the media contained only 25 mg/liter of glucose (160 times less than the 4 g/liter used here) and cultures were passed once per day, as a result spending the majority of their time in stationary phase. Significant differences in bacterial behavior arise during stationary phase [61], and the process of repeated feast and famine causes the selective force to be non-constant. This can give rise to unusual features, such as coexisting subpopulations that thrive in different phases of the daily cycle [15]. Furthermore, the Tenaillon strains evolved under poorly oxygenated conditions, given that 10 mL cultures were kept in 15 mL tubes shaken at only 100 rpm. The large difference in hypermutator prevalence between our two studies is also indicative of the methodological influence on genomic changes; at least 20% of our strains became hypermutators (discounting those likely caused by population cross-mixing), while Tenaillon evolved 115 lines of which only one became a hypermutator, and was precluded from further analysis.

As for the carbon source availability, in this study mutations facilitating faster glucose uptake are selected for only to the extent that they enable faster growth, whereas in LTEE conditions faster uptake of the scarce glucose in and of itself provides a significant advantage. This may be why under LTEE conditions *pykF* is known to acquire an internal insertion sequence [58], almost certainly leading to full gene inactivation, but the *pykF* mutations identified herein potentially preserve some of the enzyme's functionality given the failure of off-target *pykF* mutations to accrue in the MAGE strains. PykF is one of two enzymes catalyzing conversion of PEP to pyruvate in

glycolysis, and it is hypothesized that PykF inactivation enables faster glucose uptake by increasing intracellular concentrations of PEP, which can be used by the phosphotransferase system to drive glucose uptake [33]. In the glucose-rich conditions of this study, PykF impairment as opposed to inactivation could serve as a more beneficial alternative, balancing the pros of increased glucose uptake against the cons of a decrease in glycolytic flux capabilities. Another relevant comparison exists between this study and one with identical experimental conditions and starting strain, but with the culturing temperature set to 37°C instead of 42°C [5]. The most notable genetic feature of the lower-temperature ALE was the various *rpoB* SNPs acquired by every evolved strain, but *rpoB* did not mutate in any of our higher-temperature ALE strains. Recurring mutational agreements exist only for *pyrE/rph*, *hns/tdk*, and *pykF* (which similarly experienced no alterations causing clear-cut inactivation). Together these mutational comparisons with other studies highlight the significant influence of experimental conditions and starting genotype on the resulting genetic adaptation. Thus, mutational reproducibility largely cannot be expected if even slightly different strains are used, if experimental methodology varies, or if evolution to multiple factors is simultaneous instead of sequential (e.g., evolving on glucose at 42°C vs. evolving first on glucose at 37°C, and subsequently at 42°C).

Despite notable recurrence of multiple mutations across the evolved strains of this study, other than *pyrE/rph* no gene was mutated in more than half of the endpoints. In spite of these mostly distinct genotypes, convergence occurred toward the same physiological and transcriptional state. Growth, glucose uptake, and acetate production rates increased uniformly across all strains, while biomass yield decreased. Many genes

were significantly differentially expressed in the wild-type upon an upshift from 37°C to 42°C, but following evolution at the elevated temperature more than 70% of these genes reverted their expression state back towards that of the 'unperturbed' wild-type at 37°C. Mimicking the uniform nature of the physiological changes, these expression shifts were generally highly parallel – over 700 genes shifted back towards the unperturbed state in 8 or more of the 10 evolved strains. Though a minority, highly parallel shifts that moved expression further away from the unperturbed state occurred for over 100 genes. These trends in frequency and direction of shifts are in good agreement with two previous studies, demonstrating a consistent method of adaptation whether for growth on new carbon sources [19], acclimation after perturbation of a metabolic pathway [6], or adjustment to a global stress such as elevated temperature. Any changes in physiology and gene expression observed between the wild-type and evolved strains must arise from the genetic differences that developed, thus the lack of uniformly-occurring mutations means that the different genotypes acquired by each strain result in roughly the same overall state.

That large-scale alterations to the expression state can result from relatively few genetic changes points to the influence of regulatory mutations, a category into which many of the most frequently mutated genes fell, most notably *rpoC*, *rne*, and *ygaH/mprA*. At least one of these genes was mutated in every evolved strain, a prevalence matched or exceeded only by mutations in *pyrE/rph* and *pykF*. Interestingly, these latter two genes were the only metabolically-related mutations to occur in more than 2 of the 10 evolved strains, and neither is thought to result specifically from the increased temperature of the evolution, as discussed above; temperature-specific adaptation for the most part seems to

involve changes in gene regulation or the cell envelope. The regulatory mutations clearly have a net positive benefit on the cell given their frequent occurrence, but this does not mean that every one of the widespread expression levels altered as a result is likewise beneficial. It could be that some expression shifts caused by the regulatory mutations are actively detrimental to the cell, just less so than the cumulative benefit of the global change. These deleterious changes would then be targets for further adaptation that could alleviate their effects. Here, as in previous works [62], we find several pieces of evidence supporting this hypothesis, most strikingly in the massive and counterintuitive up-regulation of flagellar genes following evolution, a trend that is resisted by two strains that succeeded in acquiring a mutation to mitigate the likely detrimental up-regulation. Although overall fitness is always selected for, it seems that over the course of an evolution certain subsystems of the cell may experience transient fitness decreases before additional mutations can develop to undo them.

Overall, the results of this study yield lessons important for the future implementation of ALE as a tool for both biological discovery and engineering. Firstly, we have demonstrated the utility of MAGE as a method by which to probe mutational causality, diminishing the need for the often-laborious creation of single and combinatorial knock-ins. Such a method would prove particularly useful in ALE studies lacking in replicates with which to examine mutational parallelism, or when hypermutability has produced an overabundance of genetic changes that obscure the strongly causal mutations hidden among them. Secondly, by comparing the mutations identified herein with those found in related studies, we have highlighted the great extent to which mutational development is predicated upon the ancestral genotype and

influenced by all aspects of the evolutionary environment. Rarely will two studies examine evolution of identical strains under identical experimental conditions, thus authors should be wary of basing their expectations on comparisons with similar works. For example, if a more heat-tolerant production strain for a particular biochemical were desired, knocking-in the mutations identified in this study would not necessarily provide a benefit in the dissimilar genetic background and culturing environment. Finally, the previously documented restoration of expression state towards wild-type levels following metabolic perturbations and subsequent evolution has been shown to extend to global stresses, namely elevated temperature, as the perturbation. Transcriptomic analysis of our evolved strains demonstrated this, as well as revealing the apparent occurrence of likely transient, localized, detrimental changes in gene expression. In much the same way that overall entropy maximization can be achieved despite entropic decreases in certain components, as with a protein folding in solution [63], genetic subsystems can experience fitness decreases in the cellular pursuit of overall fitness maximization. Although given enough evolutionary time such sub-optimal components would be ameliorated by compensatory mutations, this nevertheless provides an avenue for rational design to improve on the evolution process. Genome-scale metabolic models are capable of making *a priori* predictions on genes necessary for optimal cellular growth in a particular environment [64], so those genes predicted to be useless could be knocked out ahead of time, saving the cell from potentially energetically-wasteful expression. Starting with a 'pre-adapted' strain such as this would expedite the evolution process and eliminate the need for mutations solely to counteract superfluous expression.

## 2.4 Materials and Methods

### 2.4.1 Adaptive Evolution

An automated system was used to propagate the evolving populations over the course of the ALE and monitor their growth rates. Flasks filled with 25 mL of 4 g/L M9 minimal media were kept at 42°C through placement in a heat block and aerated by magnetic tumble stirrers at 1800 RPM. At the start of the experiment, a flask of the wild-type strain *E. coli* K-12 MG1655 (ATCC47076) was grown up to stationary phase in the same conditions and used to inoculate 10 independent flasks with 900 μL of culture. As the bacteria grew, the automated system took optical density (OD) measurements at 600 nm for each flask at four timepoints, targeted to evenly span an OD range of 0.05 - 0.3 based on the most recently calculated growth rate and the starting OD of the flask. Growth rates were determined by taking the slope of a least-squares linear regression line fit to the logarithm of the OD measurements. Once reaching the target OD of 0.3, 10 μL of culture was passed into a new flask, and in the even numbered experiments this passage volume was changed to 100 μL after 20 days of evolution. At the OD of 0.3, glucose concentration only dropped from 4 g/L to ~3.5 g/L (determined by HPLC measurement of the cultures), so exponential growth in excess glucose conditions was constantly maintained.

Growth rates for each flask were discarded as untrustworthy if fewer than four OD points were sampled, if the points spanned a range of fewer than 0.2 or more than 0.4 OD units, or if the $R^2$ correlation was below 0.99. To reduce noise, the data were smoothed by averaging each point with its five adjacent neighbors on either side after applying weights following a normal distribution ($\sigma = 2$) centered on the point in

question. The evolution trajectory curves were obtained by fitting a monotonically increasing piecewise cubic spline to the smoothed data. Fitting to the unsmoothed data resulted in negligible changes to the spline. The cumulative number of cell divisions (CCD) was calculated as outlined previously [21].

Glucose M9 minimal media consisted of 4 g/L Glucose, 0.1mM $CaCl_2$, 2.0mM $MgSO_4$, 1X Trace Element Solution and 1X 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.

### 2.4.2 Cross-mixing Analysis

The sequencing results of the evolved endpoint strains highlighted the fact that there was unintentional cross-mixing between the populations over the course of the ALE. Strains 6 (41 total mutations) and 8 (55 total mutations) shared 24 identical mutations which did not occur in any of the other isolated strains, including the SNP which caused a truncated form of MutD (DnaQ) likely responsible for the observed hypermutator phenotype [27]. Although less immediately apparent, it seems likely that strains 2 (34 total mutations) and 3 (30 total mutations) suffered from the same cross-mixing – they share 3 identical mutations (a 1,222 bp deletion, a 1 bp deletion, and a SNP) which don't occur in any of the other strains, including the *mutL* mutation which likely causes their hypermutator phenotype [26]. For these reasons, we did not consider

the mutations shared only between strains 6 & 8 or 2 & 3 to have arisen independently when performing mutational recurrence analysis.

Given these occurrences of cross-mixing it was important to establish that the non-mutator strains did not likewise share a partially-evolved ancestor, which would complicate the determination of recurring mutations. If it were the case that all identical mutations resulted from cross-mixing and did not arise independently, then we would expect to see a number of identical mutations in two or more strains, but these strains would not share identical mutations with any other strains. This was decidedly not the case (supplementary fig. S5). For example, consider the identical *rpoC* mutation shared between strains 5, 7, and 10. If we posit that this is the result of the same *rpoC* mutant invading and fixing within all three populations, then it must be that the remaining shared mutations all occurred independently (*hfq* in strains 1 & 7, *ygaH/mprA* in strains 4 & 5, and *pykF* & *mlaE* in strains 4, 5, and 7) given that the other strains do not share the *rpoC* mutation. Thus in the 'worst case' scenario, only one of these four sets of shared mutations can be explained away by cross-mixing.

### 2.4.3 DNA Sequencing

After colonies were isolated and selected on LB agar plates, genomic DNA was extracted 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 a Nextera XT kit from Illumina (San Diego, CA) with 1 ng of input DNA total. Sequences were obtained using an Illumina Miseq with a PE500v2 kit. The breseq pipeline [58]

version 0.22 with bowtie2 [65] was used to map sequencing reads and identify mutations relative to the *E. Coli* K12 MG1655 genome (NCBI accession NC_000913.2). All samples had an average mapped coverage of at least 25x.

### 2.4.4 RNA Sequencing

RNA-sequencing data were generated under conditions of exponential-phase, aerobic growth in glucose M9 minimal media. 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 Readylyse 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, 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 [66]. The rRNA was isolated using Epicentre's Ribo-Zero rRNA removal kit for Gram Negative Bacteria. Sequences were run on an Illumina Miseq using a PE50v2 kit. Reads were mapped to the *E. coli* K12 Genome (NC_000913.2) using bowtie2, allowing for up to 2 mismatches and enforcing paired-end constraints. Differentially expressed genes were determined by cuffdiff [67] with upper-quartile normalization and setting a maximum false discovery rate of 0.05. Principal component analysis was performed on the RNA-seq FPKM values using the *pca* function in MATLAB.

### 2.4.5 Physiological Measurements

Growth curves for selected clones were started from stationary phase overnight cultures and grown under conditions identical to the ALE experiment, but with sampling occurring every 20 minutes. At each sampling time the $OD_{600}$ was taken along with a small volume of the growing culture that was then filter-sterilized. The filtrate was injected into a high performance liquid chromatography (HPLC) column (Aminex HPX-87H Column #125-0140). Compound concentrations at each time point were determined by comparison to a standard curve of know concentrations, and were used to determine rates of glucose uptake and acetate secretion by the cells. No other compounds were detected in the filtrate.

### 2.4.6 MAGE

For recombineering, cells containing the pMA7 plasmid (manuscript under preparation) were grown in 15 mL LB media, with shaking at 37°C to an $OD_{600}$ of 0.4. Recombineering was mediated by the lambda Red beta protein and induced through the ParaBAD promoter for 10 minutes by adding arabinose to a final concentration of 0.2%. After induction, cells were placed on ice for at least 15 minutes before being harvested, washed, and finally resuspended in a total volume of 200 µL ice-cold sterile water. 50 µL cells were mixed with 5 pmol oligo of each oligo and electroporated in 0.1 mm gap cuvettes; 1.8 kV, 200 Ω, 25 µF. Immediately after electroporation, 1 mL LB was added to the cells. Cells were transferred to a 50 mL Falcon tube to a total volume of 5 mL LB and grown for at least 3 hours at 37°C to allow full segregation of chromosomal DNA. A total of 9 consecutive rounds of recombineering were performed. A total number of 29 oligos were designed (supplementary table S2) to recreate the majority of the mutations identified in ALE lines 1, 4, 5, 7, 9 and 10. Eight recombineering experiments were

performed in parallel: six using the oligos corresponding to each of the non-mutator ALE lines, one using all 29 oligos, and one control using oligo P9 malK (table S2). Introduction of the malK oligo does not alter fitness in the glucose environments used, but makes the cells turn purple when grown on MacConkey agar maltose plates. In this control MAGE experiment, the recombinant formation frequency was found to be on average 10% by plating the recombineered population and determining the proportion of purple colonies. Recombinant formation frequency is influenced largely by oligo length [17], which, along with all other experimental variables, was kept constant for all recombineering, thus allelic biases should not be present.

Following recombineering, the mixed cell populations were cultured and serially passed for three days, under the same conditions as the ALE experiment, to select for the strains most fit for growth at 42°C. These enriched MAGE populations were then streaked on minimal M9 glucose agar plates and single colonies were isolated, screened for improved growth rates at 42°C in a microtiter plate reader, and subjected to whole genome DNA resequencing.

Chapter 2 is a reprint of a published manuscript: Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM. 2014. " Evolution of *Escherichia coli* to 42 °C and Subsequent Genetic Engineering Reveals Adaptive Mechanisms and Novel Mutations." *Molecular Biology and Evolution*, 31(10):2647-62. The dissertation author was the primary author of the paper and was responsible for the research.

# Chapter 3

# Evolution of *E. coli* on [U-$^{13}$C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology

## 3.1 Introduction

Metabolic flux analysis (MFA) has become an invaluable technique with which to probe the metabolic flux states of an organism [68, 69]. Knowledge of these intracellular fluxes is frequently used in complement with metabolic engineering, where it can be utilized to guide the rational design of genetic manipulations necessary for chemical production [70]. MFA relies on the isotopic labeling of a "tracer" compound that is passed through the reaction network of a cell, and $^{13}$C is typically used as the stable isotope of choice [71]. Inherent to $^{13}$C-MFA experiments is the assumption that the labeled compound is not metabolized differently than the unlabeled form, but the altered mass of the isotope can cause measurable kinetic isotope effects for chemical reactions. Evidentially, although $^{13}$C is only 8% more massive than $^{12}$C, plants are known to discriminate against the heavier isotope when it comes to carbon fixation [72], and recent work has indicated that neglecting kinetic isotope effects in $^{13}$C-MFA could potentially result in errors on the same scale as GC-MS measurement errors [73]. Thus, establishing

that $^{13}C$ and $^{12}C$ can be treated interchangeably when it comes to metabolic fluxes is important for the continued use of $^{13}C$-MFA as a robust experimental technique.

Adaptive laboratory evolution (ALE) involves the continuous culturing of microorganisms in a controlled setting such that natural selection for beneficial mutations will lead to cells with improved fitness. This approach can be used for a number of purposes, such as increasing cellular tolerance to some chemical, determining how a metabolic network adapts to engineered alterations, or simply optimizing growth rate on a particular substrate [2]. ALE is ideally suited to investigate the potential metabolic impact of heavier carbon because there have already been studies on evolution to unlabeled glucose [5, 8], and long term culturing serves to amplify any differences between experiments – previously, very minor differences in experimental methodology and starting strain were found to lead to vastly different mutations following ALE experiments [18, 74]. By comparing with an unlabeled glucose evolution and keeping all variables the same except for the exchange of $^{13}C$ with $^{12}C$, effects of the heavier isotope can be determined. For example, if the stronger C-C bonds caused by $^{13}C$ were indeed impacting cellular metabolism, mutations in C-C bond breaking enzymes could potentially be enriched for. Moreover, isotopically-induced growth differences, if they exist, are more likely to become apparent in the excess nutrient conditions of an ALE experiment where the only limiting factor to growth is how quickly the cells can utilize the nutrients. For these reasons, we sought to investigate the interchangeability of $^{12}C$ and $^{13}C$ glucose in *E. coli* metabolism by evolving cultures onto [U-$^{13}C$]glucose. These evolved strains were then characterized phenotypically and genetically.

## 3.2 Results and Discussion

### 3.2.1 Evolved Phenotypic Changes

Wild-type *Escherichia coli* K-12 MG1655 was used to establish six independent cultures that were serially propagated in M9 minimal medium with uniformly labeled $^{13}$C-glucose as the sole carbon source. An automated system was used [74] to perform this Adaptive Laboratory Evolution (ALE) experiment, continually tracking the cultures' growth rates and passing them to fresh media flasks before depletion of the glucose could serve as a limiting factor to growth. Thus, selection only for cells with faster exponential-phase growth rate was ensured. The fitness trajectories of the evolving cultures very closely resemble those of a previous ALE experiment by LaCroix *et al.* in which the evolution environment was the same except for the use of naturally labeled glucose (referred to as $^{12}$C-glucose herein) [5]. Noticeable jumps in the population growth rate occurred as fitter mutants emerged and gained dominance in the cultures, but over time these diminished as the trajectories asymptotically approached the "optimal growth rate" (Fig 3.1). Cultures were evolved for 40 days, corresponding with an average of 963 generations, or 2.82 x $10^{12}$ cumulative number of cell divisions (CCD) per population [21]. Using CCD as the metric of evolutionary time is more beneficial in an ALE experiment as it factors in the population subsampling that accompanies serial passage of cultures [74]. At the endpoint of the adaptive evolution, the evolved populations were on average 44% fitter (i.e. displayed higher growth rate) than the wild-type ancestor.

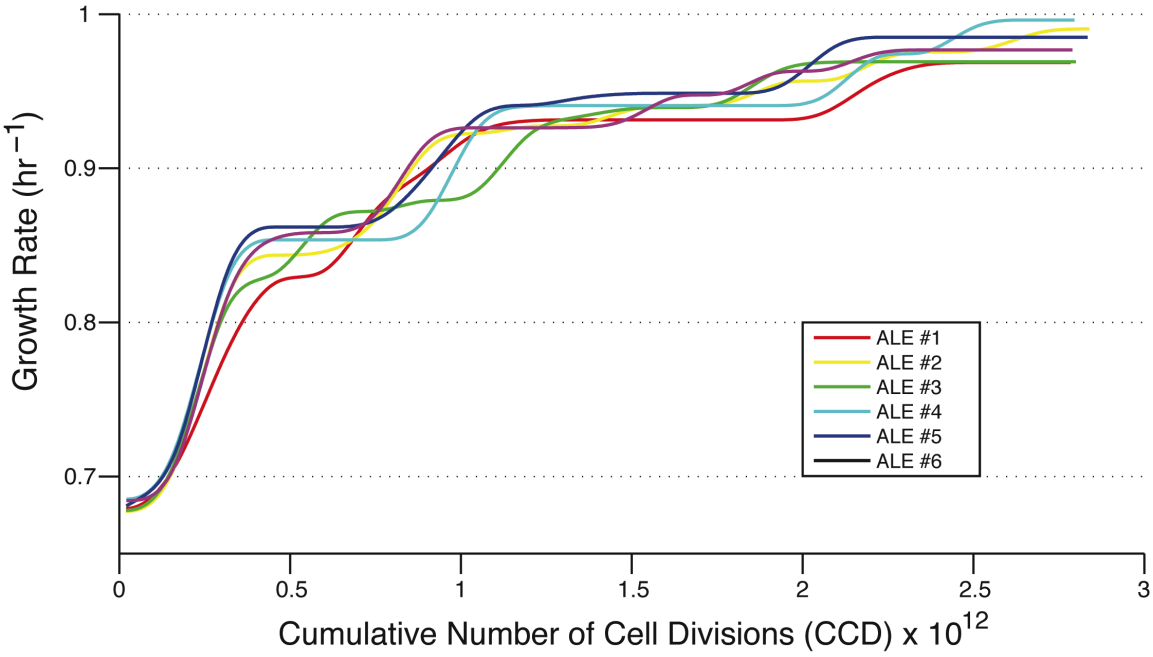



**Figure 3.1. Fitness trajectories.** Growth rates of the evolving strains over the course of the ALE as they adapted to growth on $^{13}$C glucose.

Clones were isolated from the evolved endpoint populations and subjected to phenotypic assays. Despite not having been exposed to $^{12}$C-glucose over the course of the evolution, the endpoint clones nevertheless demonstrated equivalent fitness on this substrate as they did on $^{13}$C-glucose (Fig 3.2A). Although in the wild-type strain there appears to be a slight growth advantage on $^{12}$C vs. $^{13}$C-glucose (3 ± 2 %), in the evolved strains there is no such statistically significant advantage (paired t-test, $p$ = 0.18). Again drawing a close parallel with the evolution of LaCroix *et al.,* biomass yields of both $^{12}$C-[5] and $^{13}$C-evolved strains slightly decreased, glucose uptake rates increased, and acetate production rates spanned a wide range of increases, indicating that the cells employed different metabolic tactics to realize equivalent fitness improvements (Fig 3.2B-D). Unassayed in the LaCroix study, oxygen uptake rates also increased significantly, from 49 to 92% (Fig 3.2E). Together, these phenotypic data types and evolutionary

trajectories, in comparison with equivalent data from evolution onto $^{12}$C-glucose, do not point to any differences caused by the $^{13}$C isotope.

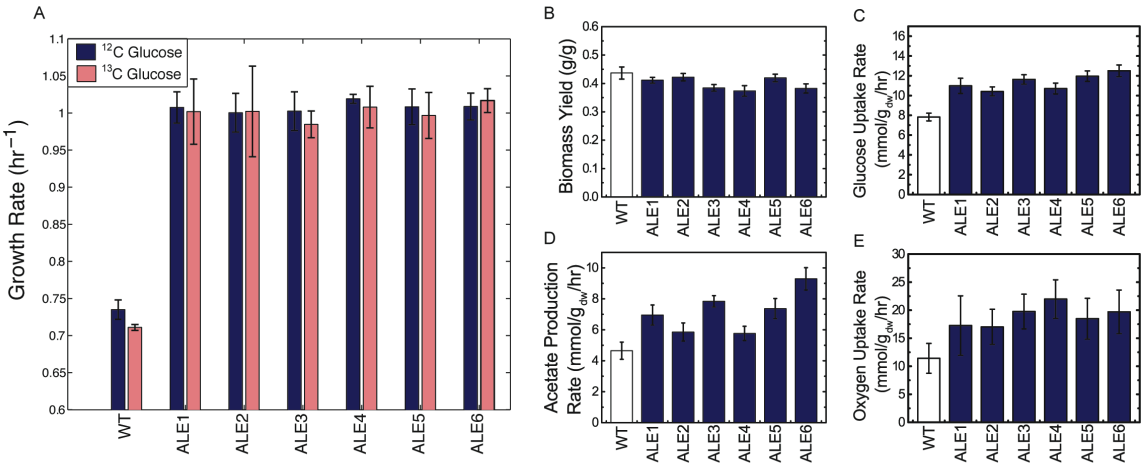



**Figure 3.2. Phenotypic characterization of the wild-type and evolved strains.** (A) Growth rates on $^{12}$C-glucose and $^{13}$C-glucose. (B) Biomass yields on $^{13}$C-glucose. (C) Glucose uptake rates on $^{13}$C-glucose. (D) Acetate production rates on $^{13}$C-glucose. (E) Oxygen uptake rates on $^{13}$C-glucose. All error bars represent standard error of the mean (n = 3).

Additionally, due to results from the genetic analyses (discussed below), it was decided to run a smaller scale ALE. This "reALE" proceeded for ten days and evolved triplicate cultures of the wild-type used previously, as well as triplicate of a strain from the LaCroix study pre-evolved on $^{12}$C-glucose. This shorter evolutionary time scale only allowed the wild-type cultures to undergo the first main jump in fitness (Fig 3.1), while the pre-evolved $^{12}$C-cultures did not appreciably change in growth rate at all (S1 Fig). Colonies isolated from the endpoints again had essentially identical growth rates on $^{12}$C and $^{13}$C-glucose (S2 Fig).

### 3.2.2 Evolved Genetic Changes

Although evolution on $^{12}C$ and $^{13}C$-glucose yields essentially equivalent phenotypes, the genetic adaptations underlying these changes could potentially differ, which would point to isotopic effects. To investigate this possibility, whole genome sequencing was performed on the phenotypically characterized endpoint clones to identify the mutations they had acquired (Table 3.1). On average, each evolved strain had about 6 mutations, with as many as 11 and as few as 3. Four genes or intergenic regions mutated in parallel across two or more of the strains. By comparing these mutations with those observed in the $^{12}C$-evolved endpoint strains [5], it was found that alterations to the same three key genetic regions (*pyrE*/*rph, rpoB*, and *hns/tdk*) appear to be responsible for the majority of fitness increases (Fig 3.3). The gene *ygaZ* mutated several times in the $^{12}C$-evolved strains but only once here, while *corA* and *iap* were not observed to mutate. However, this is unsurprising given that the $^{12}C$-adapted strains evolved for up to twice as long as this $^{13}C$ evolution (for reasons of $^{13}C$ resource conservation, and because growth rates for the $^{13}C$ ALEs had already visibly leveled off). Had the $^{13}C$ evolution proceeded for longer, it is likely that more of these smaller-effect mutations would have had time to fix in the populations, just as *ygaZ* started to.

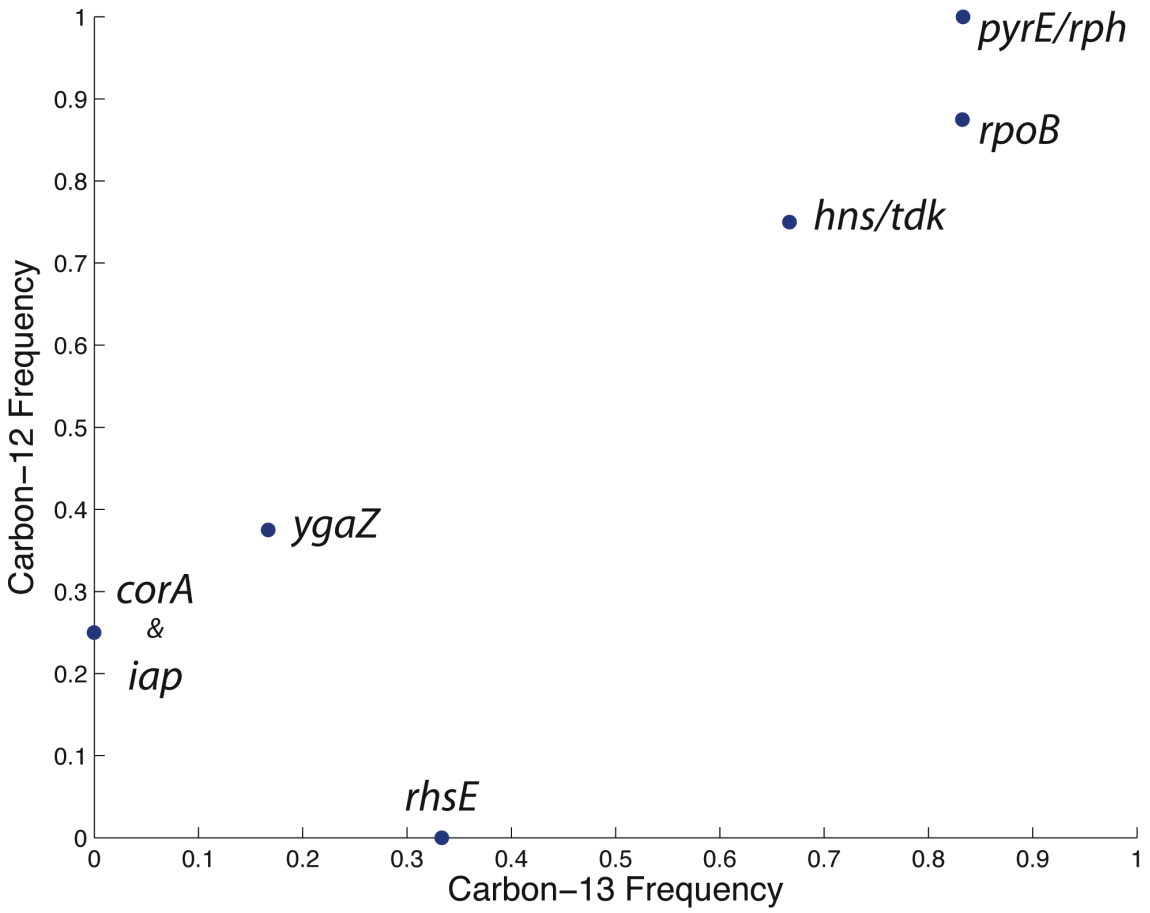



**Figure 3.3. Mutational frequency in evolved endpoint strains.** Comparison of the most frequently mutated genes observed in evolved endpoint strains grown on either $^{12}$C-glucose or $^{13}$C-glucose.

**Table 3.1. Mutations identified in the endpoints of the $^{13}$C-evolution.**

| Mutation[a] | Gene | Protein change | ALE 1 | ALE 2 | ALE 3 | ALE 4 | ALE 5 | ALE 6 |
|---|---|---|---|---|---|---|---|---|
| G→T | *murC* | V13L (GTG→TTG) | | X | | | | |
| T→G | *lolB/hemA* | intergenic (-142/-72) | | | | | | X |
| IS2 | *hns/tdk* | intergenic (-75/-526) | | | | | | X |
| IS1 | *hns/tdk* | intergenic (-112/-486) | | | X | | | |
| IS1 | *hns/tdk* | intergenic (-116/-481) | X | | | | | |
| IS5 | *hns/tdk* | intergenic (-258/-344) | | | | | X | |
| G→C | *ycjV* | pseudogene (664/957 nt) | | | | | | X |
| T→C | *rhsE* | V429A (GTT→GCT) | | X | | | | X |
| T→C | *rhsE* | F431L (TTT→CTT) | | X | | | | X |
| C→T | *rhsE* | Y432Y (TAC→TAT) | | X | | | | X |
| C→A | *rhsE* | G438G (GGC→GGA) | | X | | | | X |
| C→T | *rhsE* | L443L (CTC→CTT) | | X | | | | X |
| Δ6 bp | *pscG* | coding (778-783/939 nt) | | X | | | | |
| +GGCTGTCA | *yfjR* | coding (32/702 nt) | | | | | | X |
| T→G | *ygaY/ygaZ* | intergenic (+59/-65) | | | | X | | |
| +C | *barA* | coding (2425/2757 nt) | | X | | | | |
| T→A | *ptsP* | Q538L (CAG→CTG) | | | | X | | |
| Δ13 bp | *nirC* | coding (414-426/807 nt) | | X | | | | |
| Δ1 bp | *pyrE/rph* | intergenic (-41/+54) | | X | | | X | |
| Δ82 bp | *pyrE/rph* | intergenic | X | | X | X | | |
| +CCTGGC | *rffD* | coding (771/1263 nt) | | | | X | | |
| C→T | *rpoB* | T657I (ACC→ATC) | | X | | | | |
| G→T | *rpoB* | G1189C (GGT→TGT) | X | | X | X | X | |
| +9 bp | *rpoC* | coding (749/4224 nt) | | | | | | X |
| G→T | *mrdA* | P481T (CCT→ACT) | | | | | X | |

[a] IS = insertion sequence

Notable is the *rpoB* G1189C mutation, occurring in four of the six endpoint strains. Such an occurrence is highly unlikely unless this specific mutation is either highly causal or lays in a hotspot region for mutations, but a more plausible explanation is that it arose in the overnight culture used to inoculate the initial flask of the ALE experiment and was subsequently enriched for. The repeat appearance of the same *rhsE*

mutations suggests a similar possibility. To address these issues of genetic reproducibility the smaller scale "reALE" was started with triplicate of the wild-type and triplicate of a $^{12}$C-glucose pre-evolved strain. Of the three $^{12}$C pre-evolved reALE endpoint clones selected for sequencing, only one had any new mutations relative to its starting genotype, a 1 base pair deletion in *rpoS*. The absence of the *rpoB* G1189C mutation in the genotypes of the three wild-type reALE endpoints revealed that it had likely occurred in the starting culture of the main ALE (S2 File). Although the same *rhsE* SNPs were observed in both the main and reALE, it was subsequently discovered that these mutations were present in the purportedly-isogenic frozen wild-type stock at roughly 50% frequency. Given that *rhsE* is a pseudogene [75] whose knockout does not alter growth rate (0.73 ± 0.01 hr$^{-1}$ $^{13}$C-glucose vs. 0.73 ± 0.02 hr$^{-1}$ $^{12}$C-glucose, biological triplicates), it seems highly likely that these *rhsE* mutations are simply neutral genetic hitchhikers.

Mutations in *pyrE*/*rph* are the most frequently occurring in both $^{12}$C and $^{13}$C-glucose evolutions, and this is a well known and repeatedly seen genetic change which is thought to relieve a strain-specific defect in pyrimidine biosynthesis when grown on minimal media [28, 30]. Unique to this $^{13}$C evolution, however, is the appearance of *pyrE*/*rph* mutations other than the ubiquitous 82 base pair deletion, which has been observed multiple times in evolutions to different environments and by different labs [74, 76]. Although any frameshifting mutation that moves the *rph* stop codon closer to the *pyrE* attenuator loop appears to confer a fitness benefit [74], the 82 base pair deletion has thus far only been observed in evolution experiments because it has a large ease-of-acquisition benefit – it is flanked by two 10 base pair repeats, which allows slipped-

strand mispairing (SSM) to occur and preferentially delete the region at a much higher frequency than random SNPs or indels arise [28, 49]. Here, for the first time there is evidence for a possible decrease in the ease-of-acquisition benefit of the 82 base pair deletion, and sequencing of strains during midpoints in the evolution revealed even more atypical *pyrE*/*rph* mutations (S3 Fig and S2 File).

To investigate whether growth on $^{13}$C-glucose was acting to decrease the frequency of SSM during DNA replication, we tested an engineered strain of *E. coli* designed to allow determination of this mutational rate [77]. Added to the chromosome of this strain is a region of the *mod* gene from *H. influenzae*, which is a site of naturally occurring phase variation caused by SSM that acts on the tetranucleotide (AGTC) repeats present in the sequence, translationally fused to *lacZ*. LacZ will be produced when this construct is in-frame and colonies will appear blue when streaked on plates with X-gal, but SSM-produced insertions/deletions of the tetranucleotide repeat that cause a frameshift will lead to colonies without LacZ that will appear white on the plates. By growing the strain on $^{13}$C and $^{12}$C-glucose and calculating the rate of color switching the rate of SSM can thus be determined. However, despite suggestions to the contrary based on the genotypes of the $^{13}$C-evolved strains, the rate of SSM is not any different on $^{13}$C-glucose than it is on $^{12}$C-glucose (Table 3.2). Although no difference in SSM frequency at this tetranucleotide length scale doesn't necessarily mean there is no difference at larger (e.g., 82 base pair) scales, SSM is generally only physiologically relevant at the short scale of simple sequence repeats [78].

**Table 3.2. Measurement of slipped-strand mispairing mutation rates.**

| | Switch frequency ± SD (* $10^{-4}$) | |
|---|---|---|
| | **Blue --> White** | **White --> Blue** |
| **$^{12}$C-glucose** | 52.1 ± 5.6 | 31.0 ± 10.5 |
| **$^{13}$C-glucose** | 54.1 ± 9.2 | 36.7 ± 4.7 |
| **original study**[a] | 55.5 ± 14.6 | 22.3 ± 4.1 |

[a] Reference: [77]

### 3.2.3 Glucose Isotopic Preference

Although there are as of yet no significant indications that evolution on $^{13}$C-glucose impacted the cells differently than evolution on $^{12}$C-glucose, it remains to be seen if these evolved strains exhibit a preference for $^{13}$C, or similarly if the wild-type prefers $^{12}$C. The effects of a kinetic isotope effect would manifest in a (perhaps quite slight) difference in uptake rates of glucose based on carbon isotopic composition. This phenomenon was tested for by growing all cultures on a mixture of both $^{12}$C-glucose and $^{13}$C-glucose, and monitoring total glucose concentration and the isotopic composition over time (Fig 3.4 and S4 Fig). The calculated preference factors (f) for all strains are reported in Fig 3.4C. An f-value greater than 1 indicates preference for unlabeled glucose, while a value less than 1 reflects a preference for $^{13}$C-glucose. The only statistically significant ($p < 0.05$) preferences were small (less than 1%) preferences for $^{12}$C-glucose in ALE4 and ALE6. We suspected that these fits were affected by relatively larger GC-MS measurement errors at low glucose concentrations, and upon repeating the experiment with ALE4, ALE5, and ALE 6 we observed no statistically significant preferences (Fig 3.4C and S5 Fig). Notably, there was no evidence for a preference for $^{12}$C-glucose in the wild-type (i.e., no kinetic isotope effect), and no evidence for an evolved preference for $^{13}$C-glucose in the ALE strains. This result held for two strains,

WT and ALE1, also tested in chemostat (glucose-limited) cultures. The f-values were 1.003 ± 0.003 and 1.003 ± 0.006, respectively, demonstrating that there is negligible glucose preference across a range of conditions in which different sugar transporters may be expressed [79].

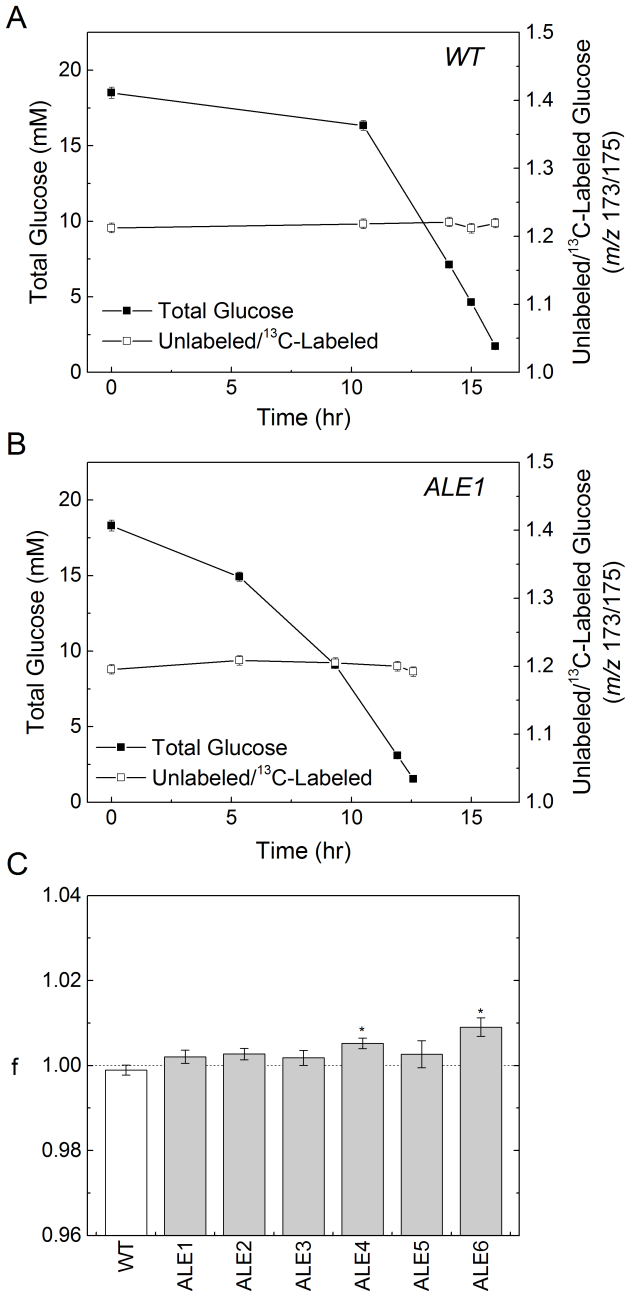



**Figure 3.4. Glucose competition.** Results of glucose competition experiments, where the medium contained a mixture of $^{12}$C-glucose and $^{13}$C-glucose. Profiles of total glucose concentration and labeling ratio are shown for the WT (A) and one representative ALE strain culture (B). The estimated glucose preference factor "f" reflects the preference for $^{12}$C-glucose vs. $^{13}$C-glucose (C). An f-value of 1 reflects no preference, greater than one reflects a preference for $^{12}$C-glucose, and less than one a preference for $^{13}$C-glucose.

## 3.3 Conclusion

In this study, six independent *E. coli* cultures were evolved for ~1000 generations on uniformly labeled $^{13}$C-glucose. The resultant strains were then characterized phenotypically and genetically. The results of these analyses, together with comparisons to an evolution study identical except for its use of unlabeled glucose [5], revealed that the heavier carbon does not significantly change any aspect of the adaptive evolution. Regardless of the isotopic composition of the glucose carbon source, populations evolving in parallel take similar trajectories across the fitness landscape. This fitness improvement is enabled by genetic changes that lead to decreased biomass yield, increased glucose and oxygen uptake rates, and increased acetate production rates. Although this trend holds uniformly, these rate and yield changes differ in their magnitude across the evolved strains despite their roughly identical growth rates. Whole genome sequencing on the evolved strains was performed and compared to the sequences of $^{12}$C-evolved strains. The only notable genetic dissimilarity was the appearance of atypical *pyrE*/*rph* mutations after $^{13}$C-evolution, despite no apparent isotopically-induced change in slipped-strand mispairing. Nevertheless, this mutational discrepancy is minor – the key genetic regions selectively mutated to optimize growth rate are identical on both $^{12}$C-glucose and $^{13}$C-glucose. The metabolic properties of the strains were also probed with a glucose competition experiment, demonstrating that the wild-type exhibits no preference for unlabeled glucose, and the evolved strains have no preference for $^{13}$C-glucose. This lack of a kinetic isotope effect is yet another indicator of the negligible influence of heavy carbon on the functioning of cells.

Overall, the data presented herein indicate that $^{12}C$ and $^{13}C$-glucose are interchangeable with respect to *E. coli* growth and metabolism, a validation of the assumption that is made ubiquitously in $^{13}C$-metabolic flux analysis experiments. This assumption is further supported by results of parallel labeling experiments, in which multiple tracers with diverse $^{13}C$ labeling are used and then simultaneously fit to a global flux solution. In an extreme recent example, Crown et. al successfully fit the results of 14 unique parallel tracer experiments in *E. coli* [80]. This would not be possible if any of the tracers caused a change in metabolism significant relative to the tight confidence intervals calculated. Thus, based on distinct data from both $^{13}C$-evolved and wild-type strains, there is strong evidence that no kinetic isotope effect can be detected in $^{13}C$-MFA studies with the current best available methods. However, careful consideration should be taken before attempting to extend these results to organisms or conditions other than the one studied herein. For example, Wasylenko and Stephanopoulos have demonstrated via mathematical modeling that isotopic effects can vary significantly across different enzymes and organisms, particularly depending on the distribution of metabolic fluxes around certain carbon-carbon bond breaking reactions, and the resulting effect on isotopic labeling can approach the magnitude of GC/MS error [73]. However, in a recent study Millard et. al demonstrated through systems-level modeling that these local effects are muted by metabolic network properties including flux control distribution and bidirectional isotope exchange [81]. The results presented here support the assertion that the kinetic isotope effect is insignificant at the physiological scale, but caution should still be applied when assessing the interchangeability of $^{12}C$ and $^{13}C$ compounds.

## 3.4 Materials and Methods

### 3.4.1 Materials

Media and chemicals were purchased from Sigma-Aldrich (St. Louis, MO). [1,2-$^{13}$C]glucose (99.5 atom% $^{13}$C), [1,6-$^{13}$C]glucose (99.5% $^{13}$C), and [U-$^{13}$C]glucose (98% $^{13}$C) were purchased from Sigma-Aldrich Isotec (St. Louis, MO). Unless otherwise noted in the text, "$^{12}$C-glucose" refers to naturally labeled glucose and "$^{13}$C-glucose" refers to [U-$^{13}$C]glucose. All experiments were performed in M9 minimal medium, which consisted of 1x M9 salts dissolved in distilled water, 2.0 mM $MgSO_4$ and 0.1 mM CaCl. Glucose was added as indicated in the text. All solutions were sterilized by filtration.

### 3.4.2 Strain and Evolution

The starting strain for evolution was wild-type *E. coli* K-12 MG1655 (ATCC 700926). For ALE, cultures were serially propagated (100 µL passage volume) in 15 mL (working volume) flasks of M9 minimal medium with 2 g/L $^{13}$C-glucose, kept at 37°C and well-mixed for full aeration. An automated system passed the cultures to fresh flasks once they had reached an $OD_{600}$ of 0.3 (Tecan Sunrise plate reader, equivalent to an $OD_{600}$ of ~1 on a traditional spectrophotometer with a 1 cm path length), a point at which nutrients were still in excess and exponential growth had not started to taper off (confirmed with growth curves and HPLC measurements). Four $OD_{600}$ measurements were taken from each flask, and the slope of $\ln(OD_{600})$ vs. time determined the culture growth rates. A cubic interpolating spline constrained to be monotonically increasing was fit to these growth rates to obtain the fitness trajectory curves.

### 3.4.3 Whole genome sequencing

Colonies were isolated from evolved populations by streaking on LB agar plates. Genomic DNA was extracted using Promega's Wizard DNA Purification Kit. The quality of DNA was assessed with UV absorbance ratios using a NanoDrop spectrophotometer. DNA was quantified using a Qubit dsDNA High Sensitivity assay. Paired-end resequencing libraries were generated using a Nextera XT kit from Illumina with 1 ng of total input DNA. Sequences were obtained using an Illumina Miseq with a PE500v2 kit. The breseq pipeline [82] version 0.23 with bowtie2 [65] was used to map sequencing reads and identify mutations relative to the *E. coli* K12 MG1655 genome (NCBI accession NC_000913.2). All samples had an average mapped coverage of at least 90x.

### 3.4.4 Slipped-strand mispairing frequency measurement

*E. coli* strain MV759 [77] was streaked onto an M9 glucose agar plate with X-gal. One white colony and one blue colony were selected and used to inoculate flasks of both $^{12}C$ and $^{13}C$-glucose M9 medium. After growing up, a volume of these cultures was plated onto $^{12}C$ or $^{13}C$-glucose M9 agar plates with X-gal, and the number of cells of each color was counted. This process was repeated independently three times, and an average of ~1200 cells were counted on each plate. The SSM-induced color switching frequency was calculated by dividing the fraction of cells that had switched color by the generations of growth that had occurred.

### 3.4.5 Glucose competition experiments

Glucose competition experiments were performed to determine the preference of the various *E. coli* strains for $^{12}$C vs. $^{13}$C-glucose. Cells were first pre-cultured overnight in medium containing 2 g/L $^{12}$C-glucose. Next, 50 μL of the overnight pre-culture was used to inoculate 10 mL of M9 medium containing approximately equal amounts (2 g/L) of $^{12}$C-glucose and $^{13}$C-glucose. Total glucose concentration was measured over time, as well as $^{13}$C-labeling of glucose remaining in the medium. Based on this data, a preference factor (f) was calculated, defined as follows:

$$\frac{uptake\ rate\ of\ unlabeled\ glucose}{uptake\ rate\ of\ labeled\ glucose} = f \times \frac{fraction\ of\ unlabeled\ glucose\ in\ medium}{fraction\ of\ labeled\ glucose\ in\ medium}$$

An f-value greater than one indicates a preference for $^{12}$C-glucose, while a value less than one reflects a preference for $^{13}$C-glucose. To determine the f-value, the following expression was used (see S4 File for derivation):

$$ln\left(\frac{x_{unlabeled}(t) * gluc(t)}{x_{unlabeled}(t=0) * gluc(t=0)}\right) = f * ln(\frac{x_{labeled}(t) * gluc(t)}{x_{labeled}(t=0) * gluc(t=0)})$$

Where, *gluc(t)* is the measured glucose concentration over time, $x_{unlabeled}(t)$ is the fraction of $^{12}$C-glucose in the medium over time, and $x_{labeled}(t)$ is the fraction of $^{13}$C-glucose in the medium over time. The f-parameter and its error estimate were obtained with the above equation by one-parameter linear least-squares regression. To reduce sensitivity to

measurement error in the initial values (t=0), the average fractional labeling of the first two samples was used as $x_{labeled}(t=0)$ and $x_{unlabeled}(t=0)$.

For the competition experiments in chemostat mode, M9 feed medium with approximately 3 mM each of $^{12}C$-glucose and $^{13}C$-glucose was prepared. The working cell culture volume was 15 mL, and the feed rate was 3 mL/hr, resulting in a dilution rate of 0.2 $hr^{-1}$. Media samples were taken after 20 hours, well after steady state was established. Error estimates of the f-value were obtained by propagating the typical GC/MS error of 0.3%.

### 3.4.6 Analytical methods

Cell growth in glucose competition experiments was monitored by measuring the optical density at 600nm ($OD_{600}$) using a spectrophotometer (Eppendorf BioPhotometer). The $OD_{600}$ values were converted to cell dry weight concentrations using a pre-determined $OD_{600}$-dry cell weight relationship for *E. coli* (1.0 $OD_{600}$ = 0.32 $g_{DW}/L$). Glucose concentration was measured with a YSI 2700 biochemistry analyzer (YSI, Yellow Springs, OH). Acetate concentration was measured by HPLC [83]. No compounds other than glucose and acetate were detected in the cultures.

### 3.4.7 Determination of yields and biomass specific rates

Yield of biomass on glucose was determined as the slope of least-squares regression of biomass concentration versus glucose concentration [84]. Acetate yield was determined based on the initial and final glucose and acetate concentrations in batch cultures. Specific growth rate (GR, $h^{-1}$) was determined as the slope of least-squares

regression of ln($OD_{600}$) versus time during the exponential growth phase, typically for $OD_{600}$ values between 0.01 and 0.7. Growth rates as presented in Fig 3.2A and S2 Fig represent "physiologically adapted" rates [74], i.e. cultures were kept in exponential phase growth for three flasks so that their growth rates stabilized and were no longer decreased due to recently coming out of stationary phase [23]. This was to allow for comparison with the fitness trajectories, which inherently represent physiologically adapted rates. For calculations of compound uptake/production rates, the non-physiologically adapted (i.e. from the first post-overnight culture growth flask) growth rates were used, to allow for proper comparison with other studies (values given in S1 File). Specific glucose uptake (GUR, mmol/$g_{DW}$/h) was determined as the ratio of growth rate to biomass yield. Specific acetate production rate (APR, mmol/$g_{DW}$/h) was determined as the product of specific glucose uptake rate and acetate yield. Specific oxygen uptake (OUR, mmol/$g_{DW}$/h) was determined based on electron balance as described previously, assuming a degree of reduction of 4.35 and a molecular weight of 0.0255 $g_{DW}$/mmol-C for dry biomass [85]:

$$\text{OUR (mmol/}g_{DW}\text{/h)} = (24\times\text{GUR} - 8\times\text{APR} - 4.35/0.0255\times\text{GR}) / 4$$

### 3.4.8 Gas chromatography mass spectrometry

GC-MS analysis was performed on an Agilent 7890B GC system equipped with a DB-5MS capillary column (30 m, 0.25 mm i.d., 0.25 μm-phase thickness; Agilent J&W Scientific), connected to an Agilent 5977A Mass Spectrometer operating under ionization by electron impact (EI) at 70 eV [86]. Labeling of glucose was determined by GC-MS analysis of the aldonitrile pentapropionate derivative of glucose [87]. The fragments at

*m/*z 173 and *m/*z 370 were analyzed, which contain the last two and first five carbon atoms of glucose, respectively. Mass isotopomer distributions were obtained by integration and corrected for natural isotope abundances [88].

Chapter 3 is a reprint of a published manuscript: 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(3):e0151130. The dissertation author was the primary author of the paper and was responsible for the research.

# Chapter 4

# Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies

## 4.1 Introduction

In heterotrophs such as *E. coli*, catabolism of carbon substrates is the driving force behind the energy generation and chemical synthesis necessary for homeostasis and anabolism [89]. Although glucose is the most readily metabolized carbohydrate [90], the frequent environmental availability of other carbon compounds long ago led most organisms to evolve the ability to subsist on a range of nutritional sources. These alternative compounds can vary greatly in energy content, the point at which they enter into the metabolic network, and impact on cellular phenotype [91]. Alternative compounds' ability to sustain growth and proliferation in the absence of glucose makes it almost essential for a robust bacterium to be able to switch between carbon growth substrates as environmental circumstances dictate. For example, an enteric *E. coli* bacterium depends on its ability to switch to metabolizing different carbohydrates as it

passes through the digestive tract [92]. Such switching between carbon sources has relevance to more than just natural environments – the use of genetically engineered microbes to produce commercially valuable chemicals frequently relies on batch growth, which can include a stage at which the cells run out of the preferred nutrient (e.g., glucose) and have to switch to a less-than-optimal alternative (e.g., xylose) [93].

Understanding diauxic shifts has been a long-standing effort of the scientific community [94]. Diauxic lag has traditionally been understood to result from catabolite repression, wherein the depletion of the preferred substrate relieves the repression on genes for metabolizing the remaining substrate(s). However, recent work has shown that cells in a multi-substrate environment can display divergent bet-hedging behaviors, which can result in subpopulations that grow differently on the substrates [95]. Moreover, slightly different microbial strains can have notably different lag-phase durations and behaviors, which can be targets for natural selection in a competitive environment [96]. Adaptive Laboratory Evolution, or ALE, serves as a technique that harnesses natural selection to arrive at genetic and phenotypic outcomes that are difficult to predict *a priori* [2]. ALE work so far has examined *E. coli* adaptation to a number of environments characterized by temporal heterogeneity – fluctuations in temperature [97, 98], pH [99], UV irradiation [100], and random stressors [101] have all been studied. It is well established that homogenous environments tend to develop narrow niche width "specialists" while heterogeneous environments usually lead to broader niche width "generalists" [102], to the extent that the failure of a generalist to develop in certain fluctuating environments is seen as surprising [103]. However, evolution of *E. coli* on a glucose/acetate mixture (of which the glucose is first depleted before a diauxic shift to

acetate occurs, creating temporal variability) has been shown to repeatedly lead to coexisting specialists rather than generalists [104], due to competition for limited resources and the fitness trade-offs of glucose vs. acetate specialization [105]. While several cases of *E. coli* evolution to alternating growth substrates have been studied, few substrates have been examined [106] and analyses have been limited to fitness assays [107] or in-depth study of a single operon [108].

In this study, we sought to investigate switching with different compounds of industrial relevance, and examine evolutionary outcomes via a number of different data types. ALE was used to adapt *E. coli* cultures for ~1000 generations to a dynamic, nutrient-excess environment in which the available carbon substrate alternated with every tube of growth, and cultures were serially propagated while still growing exponentially. Though such resource-abundant laboratory environments have few natural counterparts, the conditions of excess help to ensure that selection occurs for growth rate, without the complicating factors of resource competition or changing growth phases that introduce new stressors [109]. Populations evolved to these switching environments had substrate-specific fitnesses comparable to those reached by single-substrate-evolved control cultures. The adaptive mechanism used to achieve this fitness improvement varied based on substrate, and evolved strains likewise exhibited phenotypic, genetic, and transcriptomic dynamics that varied across compounds. Genome-scale metabolic models were used to help interpret substrate-specific results and to explain the propensity for specialist or generalist development given different switching setups.

## 4.2 Results

### 4.2.1 Evolved Population Fitness

Adaptive Laboratory Evolution (ALE) was employed to adapt *Escherichia coli* to an environment with constantly alternating carbon growth substrates. Wild-type *E. coli* K-12 MG1655 was used, and substrate switching was examined on four different carbon compounds: glucose, xylose, glycerol, and acetate (*E. coli* wild-type growth rates at 37°C being 0.73, 0.55, 0.45, and 0.27 $hr^{-1}$, respectively). ALE replicates were each assigned a substrate in addition to glucose, and at the end of each growth tube a culture was passed to a fresh tube of M9 minimal medium containing the next carbon source such that substrate procession was glucose → [substrate] → glucose → [substrate] → etc. (Figure 4.1A). Three tubes switched between glucose/xylose (1180 generations of growth; 164 tubes of growth; 3.1 x $10^{12}$ cumulative cell divisions), three between glucose/glycerol (1170 generations; 162 tubes; 3.0 x $10^{12}$ CCD), and three between glucose/acetate (650 generations; 91 tubes; 1.8 x $10^{12}$ CCD). Because each tube underwent serial passage at the same optical density (at a point at which nutrients were still in excess and cultures were still growing exponentially), the generations/tubes/CCDs of growth upon each individual substrate were half of the total. As controls, cultures were also evolved to each of the four different compounds on their own, with no switching.

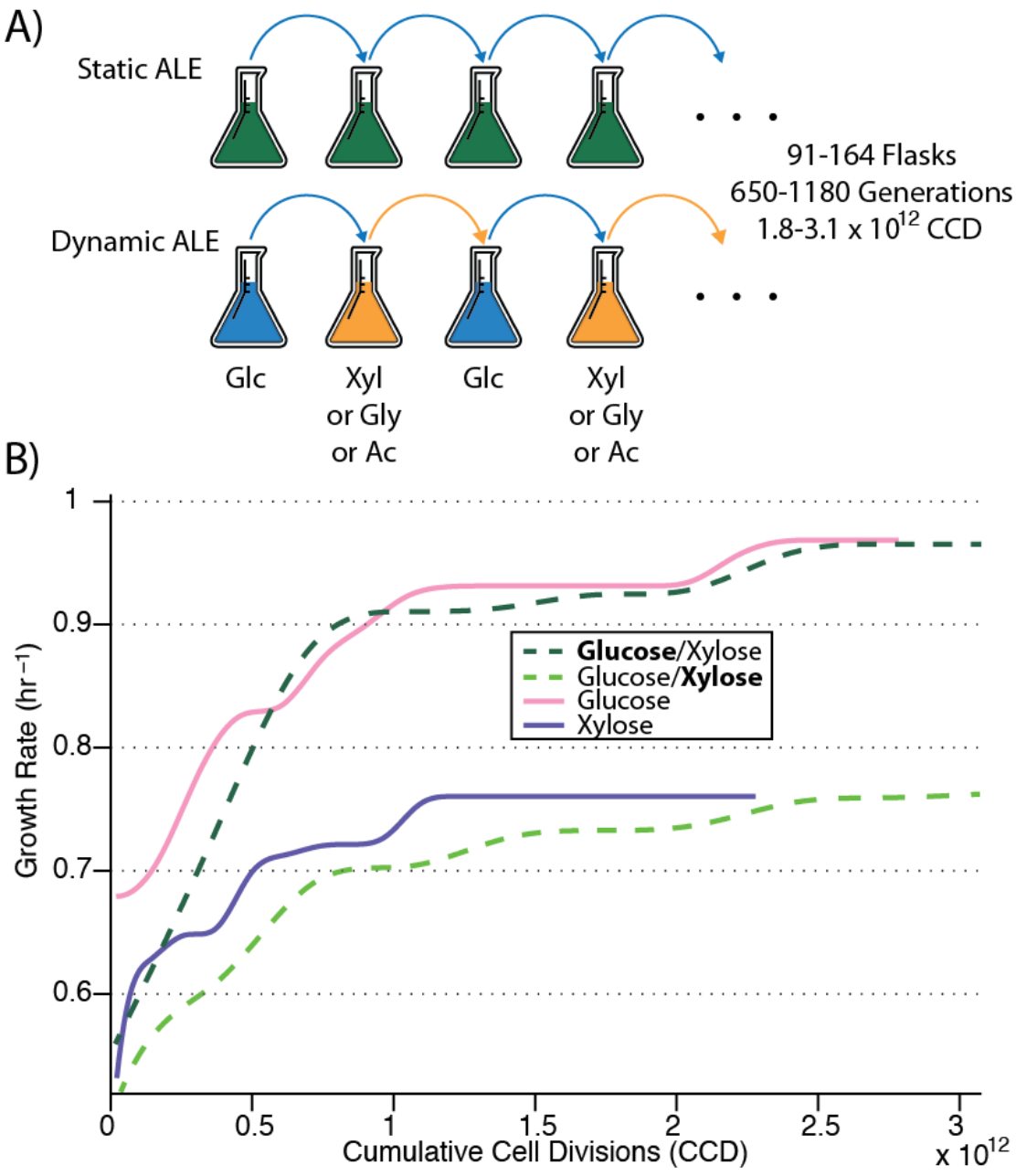



**Figure 4.1: Experimental setup and evolutionary trajectories** (A) A "static ALE" exposes cultures to a constant environment, while a "dynamic ALE" introduces temporal variability in carbon growth substrate. (B) An example plot of fitness trajectories (i.e., growth rates) for statically evolved (pink/purple) and switching (dark/light green) cultures on glucose and xylose over the course of the ALE experiment. Similar fitnesses are reached, though under switching conditions cultures take longer to get there. Glc = glucose; Xyl = xylose; Gly = glycerol; Ac = acetate

Comparison of evolutionary trajectories between the "dynamic" (switching) and "static" (non-switching) populations gives the first indication of consequences of the substrate switching. As a general trend across all compounds, replicates undergoing switching adapted at a slower rate, but were still able to reach comparable final fitness values as those control populations exposed solely to a single compound (Figure 4.1B). Although establishing a statistically significant difference between the final growth rates is made more difficult by the small number of samples, average final population growth rates on a compound failed to differentiate dynamic and static ALE setups (one-way ANOVA, $p$=0.05). However, cultures under static conditions reached their half-maximal final growth rates, on average, 43% faster than switching cultures (Supplementary File 1). Thus, it seems that adaptation proceeds at a faster rate when the selection pressure is more sustained (in this case, greater "time under selection" leads to a greater selective force), but fitness plateaus towards a similar value since the same objective is still being optimized for (namely, exponential phase growth rate on a compound).

### 4.2.2 Genetic Analysis

Population sequencing was performed on the cultures evolved under dynamic conditions to examine whether specialists (coexisting subpopulations optimized for the different substrates) or generalists (one main dominant strain capable of good growth on both substrates) had developed. The endpoint populations for each dynamic ALE replicate were grown up on each of the different carbon sources they had been switching between and population sequencing results were compared (see Methods). Consider, as an example, a population switching between glucose and acetate. If specialist

subpopulations are present, then growing the culture on glucose will preferentially enrich for the glucose-specialist, and likewise for the acetate-specialist with acetate growth. Population sequencing would then show a significant change in mutational frequency dependent upon which growth enrichment was performed. If there were no specialist subpopulations, but instead one dominant generalist, then the mutational frequencies would be roughly identically regardless of substrate used for growth enrichment. Applying this sequencing analysis to all evolved replicates revealed the various compounds' propensities to elicit subpopulation development (Supplementary File 2) – all three glucose/glycerol replicates were isogenic, all three glucose/acetate replicates were subpopulations, and two of three glucose/xylose replicates were clearly isogenic, while the third replicate developed hypermutability which complicated the genetic analysis. As examples for each case, the glucose/acetate switcher Glc/Ac #3 developed two distinct specialist populations (a *sapA/rpoC* mutant for glucose growth and a *nrd/ptsP/rpoC* mutant for acetate growth, on top of two mutations which swept both subpopulations), while the glucose/glycerol switcher Glc/Gly #2 was a generalist, having the same population composition regardless of growth substrate (Table 4.1).

**Table 4.1: Substrate-induced mutational frequency differences or lack thereof**

**Apparent Specialists (Glc/Ac #3)**

| | | population frequency (%) | |
|---|---|---|---|
| **mutation** | **gene** | **Glucose** | **Acetate** |
| L14R (CTT→CGT) | *sapA* | 85 | 0 |
| Δ88 bp | *nrdA/nrdB* | 0 | 68 |
| Δ1 bp (708/2247 nt) | *ptsP* | 0 | 100 |
| R98H (CGC→CAC) | *rpoC* | 95 | 11 |
| K398M (AAG→ATG) | *rpoC* | 7 | 79 |
| IS5 (141-144/144 nt) | *yobF* | 100 | 100 |
| Δ1 bp intergenic (-33/+33) | *pyrE/rph* | 100 | 100 |

**Apparent Generalist (Glc/Gly #2)**

| | | population frequency (%) | |
|---|---|---|---|
| **mutation** | **gene** | **Glucose** | **Glycerol** |
| IS5 (258-261/491 nt) | *ybbD* | 35 | 42 |
| Δ82 bp | *pyrE/rph* | 84 | 90 |
| M272I (ATG→ATT) | *glpK* | 100 | 100 |
| E672K (GAA→AAA) | *rpoB* | 100 | 100 |

Comparing mutational frequency across the switching-adapted cultures and single-substrate-evolved controls revealed key genes under selective pressure in these conditions (Table 4.2). Many of the repeatedly mutated genes are seen frequently in other evolution experiments, with the fitness benefit already either known or inferred – *pyrE*/*rph* mutations for improved minimal media growth [28], *pykF* and *hns*/*tdk* mutations for glucose growth [5, 74], *glpK* mutations for glycerol growth [7], and *rpoB* and *rpoC* mutations to serve as large scale transcriptional re-wirings [36, 110]. In addition to these oft seen mutational targets, several genes stood out as indicators of differing adaptive strategies between dynamic and static conditions. Most striking are mutations in the *cspC/yobF* region, in which a number of distinct mutations were seen across five out of six glucose/xylose and glucose/acetate populations (with the sixth having a mutated *cspE* rather than *cspC*), but only in a single acetate-evolved replicate.

These dynamically-favored mutations may be altering CspC's role as a transcript stabilizer in stressful environments, a functionality more relevant as a target for adaptation when the carbon source is frequently changing [111]. Similarly, mutations in *ptsP* (3 unique dynamic mutations, with 2 in one lineage), *relA* (3 unique dynamic mutations), and *sapB* (2 unique dynamic mutations) appeared to be dynamically-favored, while mutations in *rho* (4 unique static mutations) and *xylR* (2 unique static mutations) appeared statically-favored. Although, as a whole, these data indicate differing adaptive strategies for dynamic and static growth environments, the explicit biochemical mechanisms through which such mutations enable fitness improvements remain unclear without detailed follow-up analyses [7, 110].

**Table 4.2: Genetic regions mutated repeatedly during selection**

| Gene | Unique Mutations | Dynamic Occurrences | Static Occurrences |
|---|---|---|---|
| *rpoC* | 8 | Glc/Xyl (3/3)<br>Glc/Gly (1/3)<br>Glc/Ac (3/3) | Glc (1/3)<br>Xyl (2/4)<br>Gly (2/2) |
| *cspC/yobF* | 6 | Glc/Xyl (2/3)<br>Glc/Ac (3/3) | Ac (1/3) |
| *rho* | 5 | Glc/Ac (1/3) | Xyl (2/4)<br>Ac (2/3) |
| *glpK* | 5 | Glc/Gly (3/3)<br>Glc/Xyl (1/3) | Gly (2/2) |
| *ptsP* | 4 | Glc/Ac (2/3) | Ac (1/3) |
| *rpoB* | 4 | Glc/Gly (2/3)<br>Glc/Ac (1/3) | Glc (2/3) |
| *pyrE/rph* | 4 | Glc/Gly (2/3)<br>Glc/Ac (3/3) | Glc (2/3)<br>Xyl (4/4)<br>Ac (1/3) |
| *relA* | 3 | Glc/Xyl (3/3) | n/a |
| *pykF* | 3 | Glc/Xyl (1/3)<br>Glc/Ac (2/3) | n/a |
| *xylR* | 2 | n/a | Xyl (2/4) |
| *sapB* | 2 | Glc/Xyl (1/3)<br>Glc/Ac (1/3) | n/a |
| *hns/tdk* | 2 | n/a | Glc (2/3) |

### 4.2.3 Physiological Analysis of Evolved Strains

Representative clones were isolated from evolved cultures for purposes of physiological characterization. For cultures that evolved substrate specialists, this involved isolating a clone for each of the observed subpopulations. Clones were sequenced to ensure that they were representative of the evolved cultures, containing the same key mutations revealed from population sequencing (Supplementary File 2). The substrate switching phenotypes of the clones were characterized via a series of diauxic growth curves. Clones were grown in the presence of 5 mM glucose and 5 mM of an additional carbon substrate, depending on the environment the clone had been exposed to over the course of the evolution. The 5 mM concentrations were chosen such that sufficiently dense final Optical Densities (ODs) could be obtained, but not without utilizing both of the available carbon compounds.

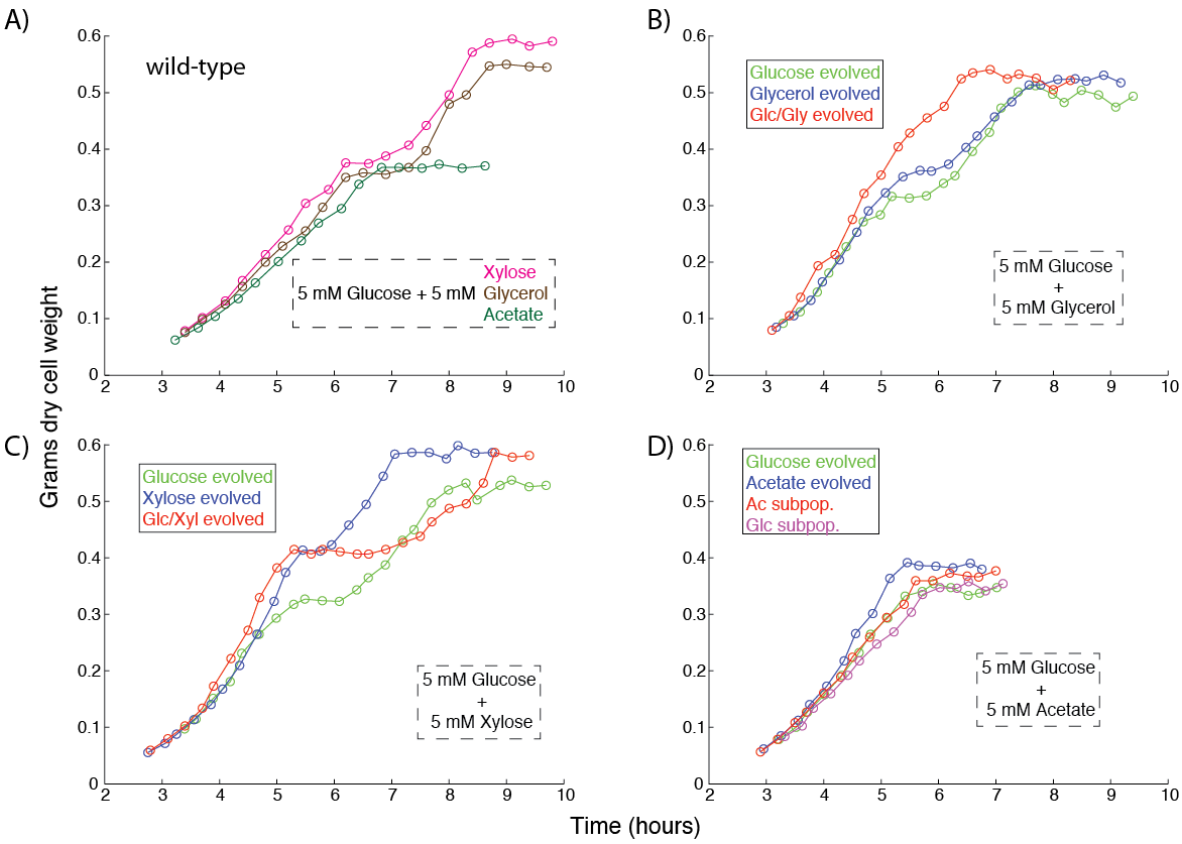



**Figure 4.2: Diauxic growth curves** (A) Wild-type performance on all substrate mixtures. (B) Evolved strains on glucose + glycerol. (C) Evolved strains on glucose + xylose. (D) Evolved strains on glucose + acetate.

Unsurprisingly, the different carbon substrates led to different growth phenotypes for the various evolved strains (Figure 4.2). The wild-type achieved the highest density on glucose/xylose, an intermediate density on glucose/glycerol, and the lowest density on glucose/acetate. While xylose and glycerol both caused diauxic lag phases, there was no such lag when the wild-type was grown on glc/ac, likely due to acetate's frequent presence as an overflow metabolite in regular batch culturing [112] and inability to trigger anabolism following glucose exhaustion [113]. Nevertheless, overall growth rate on glc/ac improved in all evolved strains, with the static acetate-evolved strain growing most robustly. Similarly, the static xylose-evolved strain outperformed the other strains under conditions of glc/xyl diauxie. The dynamically-evolved glc/xyl generalist's

*increased* lag phase duration appears counterintuitive, but a simultaneous multi-substrate environment is not something the cells were ever exposed to during the ALE – in this case inter-tube substrate switching adaptation did not extend to intra-tube switching. Contrastingly, the clone evolved to an environment switching between glucose and glycerol was able to completely abolish the lag phase that typically occurs midway through growth on both substrates. Moreover, this lag phase was not abolished in either of the statically-evolved glucose or glycerol controls, leading to much different performance in the diauxic growth test. Overall, the evolved strains' variable diauxic growth phenotypes across substrate conditions highlight the complexity of adaptation to dynamic environments.

### 4.2.4 Transcriptomic Analysis of Evolved Strains

RNAseq was performed on both statically- and dynamically-evolved clones to probe the transcriptional states of the strains under relevant substrate growth conditions. For a given growth environment, principal component analysis was expected to cluster statically-evolved strains together in a region corresponding to the "optimal" expression state for fast growth on that substrate. Dynamically-evolved strains, however, would be expected to cluster apart from the static controls given that their expression state evolved in response to multi-substrate exposure. This transcriptomic "distance" serves as an indicator of dissimilarity that should mirror substrate differences – for example, if [substrate1] is more similar to glucose than [substrate2], then a strain evolved dynamically to glucose/[substrate1] should fall closer to glucose-evolved controls than a glucose/[substrate2] strain. Indeed, PCA reinforced the conclusion drawn from

population sequencing that growth of evolved strains on glycerol and xylose create cellular states more similar to glucose than acetate – glc/xyl and glc/gly generalists clustered closer to glucose optimality than do the glc/ac specialist strains (Figure 4.3A). Similar logic implies that the dynamically evolved strains have moved closer to transcript optimality, represented by the statically-evolved controls, than the wild-type starting strain, which is what is seen for glycerol (Mahalanobis distance to static strains = 4.89 for wild-type, 2.63 for dynamic) and xylose (1.88 for wild-type, 1.35 for dynamic) (Figure 4.3B-C). However, in the case of glc/ac switching, the specialist strains appear to have adopted a different transcriptional strategy than acetate-evolved controls, falling further away from the "optimum" expression state of the static strains than does the wild-type (Figure 4.3D). This specialist discrepancy may result from the strains' nature as coexisting subpopulations and the resultant transcriptional adjustments necessary for alternating dominance between tubes during the ALE experiment. Although it is possible that the two specialists interact and exhibit different phenotypes when cultured together vs. independently, this appears unlikely given the precipitous drop in frequency of acetate-characteristic mutations after glucose enrichment, and likewise for glucose-characteristic mutations after acetate enrichment (Table 4.1, Supplementary File 2).

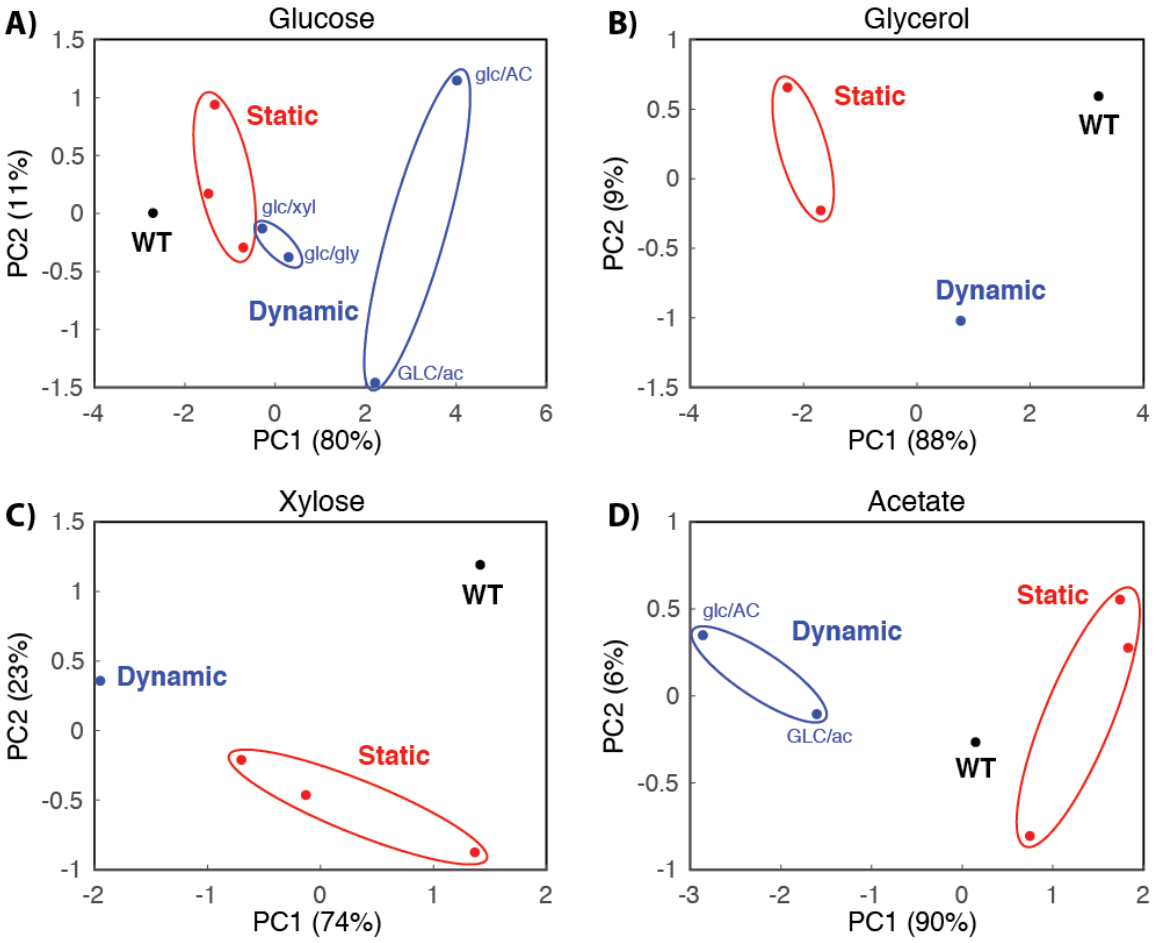



**Figure 4.3: Principal Component Analysis of RNAseq on various growth substrates** PCA plots clustering different strains' expression states grown on (A) glucose, (B) glycerol, (C) xylose, and (D) acetate. Overall, the statically evolved strains grouped together. In addition, for growth on glucose, the generalist strains grouped together closer in proximity to the static strains than did the specialist strains. glc/xyl = glucose/xylose generalist; glc/gly = glucose/glycerol generalist; GLC/ac = glucose specialist subpopulation; glc/AC = acetate specialist subpopulation; WT = wild-type

Hierarchical clustering of expression data for the statically-evolved strains further highlights the differing transcriptional strategies adopted in pursuit of substrate optimality (Figure 4.4A). Glucose and glycerol strains clustered most closely together, followed by xylose, with acetate adaptation resulting in the most distinct pattern of relative gene expression levels. For dynamically-evolved strains, these substrate-optimal transcriptional patterns cannot be adopted without leaving the cells in a state where

widespread expression changes must be made between every growth tube, an adaptation strategy unlikely to prove optimal given the time and energy it takes to alter expression levels [114]. It would be expected that for the rapidly switching ALE environment utilized herein, the transcriptome will adapt such that shifts in expression level needed between the two relevant substrates become minimized. To test this, RNAseq data for individual strains (wild-type or evolved) was compared to determine how expression levels changed when a strain was grown on glucose vs. when it was grown on [alternative substrate] (Figure 4.4B-C). For glucose/glycerol switching, the wild-type starting strain had 1689 genes with greater than 2-fold changes in gene expression across the two substrates and an average gene fold change magnitude of 6.9, while the evolved glucose/glycerol generalist significantly decreased (paired t-test, $p$=0.018) to 665 genes and a 2.3-fold average expression change, respectively (Figure 4.4B). Similarly, the wild-type on glucose/xylose had 1676 genes with |fold change|>2 and an average fold change of 4.9, which significantly decreased ($p$=0.034) in the glucose/xylose generalist to 842 genes and a 2.9-fold average change. This stands in stark contrast with the phenotype of the dynamically evolved glucose/acetate specialist strains, which did not approach any closer to transcriptional parity across the substrates than the wild-type strain (Figure 4.4C). Unlike the generalists, the coexisting glucose and acetate specialist subpopulations obviated the need for a single genotype capable of reconciling the dissimilar optimal expression states between the two substrates. Thus, it becomes clear that global transcriptional analyses across substrates can be used to interpret the differential development of specialists or generalists under various substrate switching regimes.

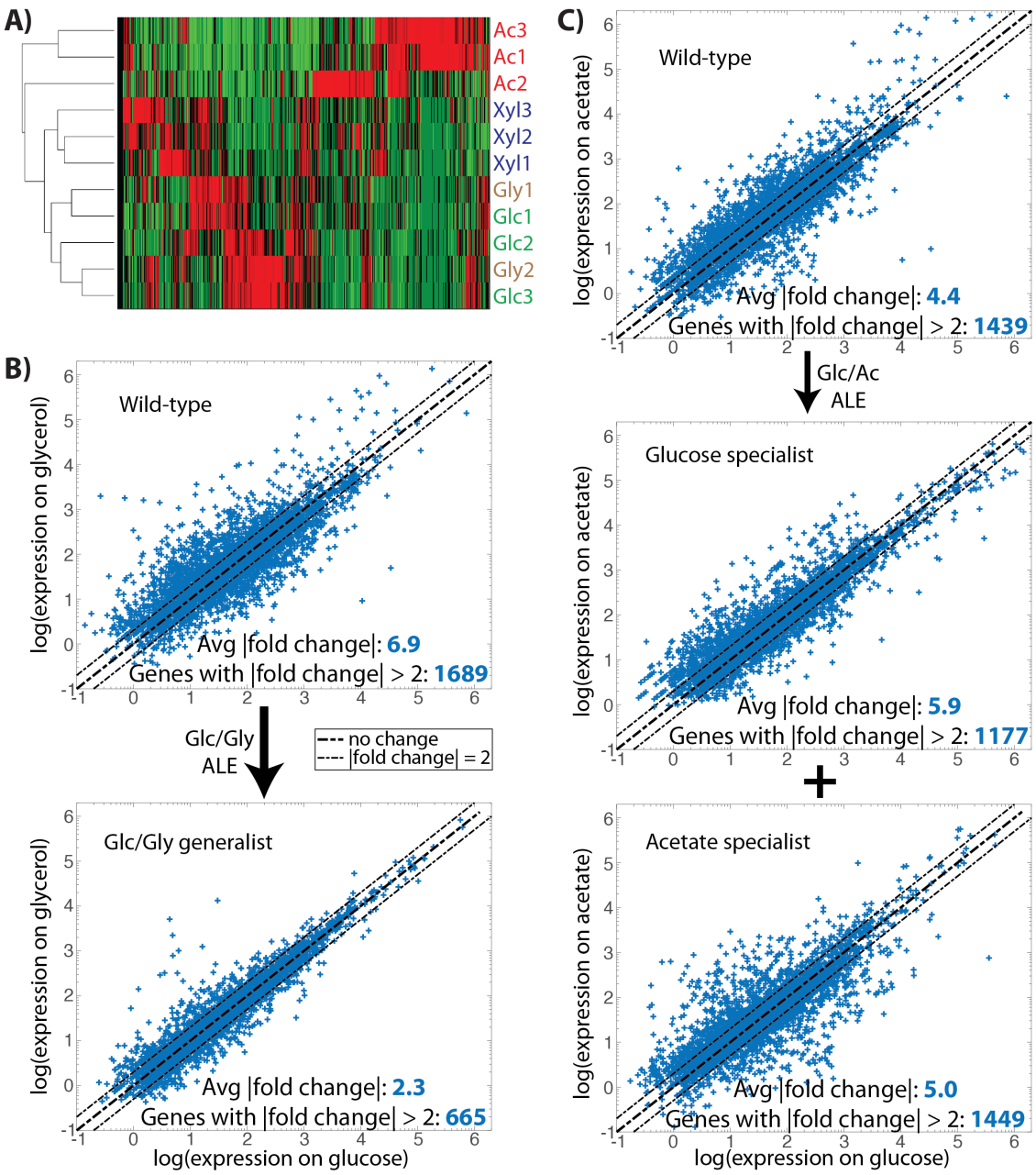



**Figure 4.4: Global transcriptome changes in statically and dynamically evolved strains** (A) Hierarchical clustering of strains statically evolved to each of the four studied substrates. Strains optimized for glucose and glycerol growth have the most similarity in expression state, followed by xylose and then acetate. (B & C) Scatter plots of expression data for the wild-type ancestor and dynamically evolved strains when grown on (B) glucose vs glycerol or (C) glucose vs acetate. Points falling on the diagonal represent genes whose expression does not vary, while those falling outside the dashed lines are genes with more than 2-fold changes in expression across the two substrates. (B) The number and magnitude of transcriptional shifts between glucose and glycerol growth conditions significantly decreased in the dynamically evolved generalist strain. (C) The number and magnitude of transcriptional shifts between glucose and acetate growth conditions did not significantly change in the dynamically evolved specialist strains.

### 4.2.5 Metabolic Modeling

To better understand the mechanisms underlying growth substrate-driven adaptive responses, genome-scale metabolic modeling techniques were applied. By modeling the metabolic reaction network of an organism with a stoichiometric matrix [115] and applying relevant physiological constraints, predictions can be made on optimal growth behavior and metabolic flux states in a particular environment [116]. Here, the relevant environments are identical except for the different carbon growth substrates.

The experimental analyses performed thus far indicate that ALE cultures adapted differently to different switching schemes, becoming either generalists or specialists in response to different substrate combinations. Robust growth under glucose/xylose and glucose/glycerol switching was achieved by versatile generalist strains, whereas glucose/acetate switching failed to select for such strains, instead opting for co-existing specialist subpopulations as the strategy for growth improvement. One potential explanation for this discrepancy is that glucose and acetate are too metabolically dissimilar (requiring conflicting or disparate pathways for metabolism) for a single strain to easily evolve for robust growth on both, while xylose and glycerol are both similar enough to glucose to avoid this. To test this, optimal growth of the wild-type strain was modeled on the four different relevant substrates individually, and the flux states necessary for optimal growth were inferred from Monte Carlo sampling of a genome-scale M-model [117]. This sampling yielded flux distributions for each chemical reaction in the model, and the three nonstandard substrates were pairwise compared with glucose (Figure 4.5). As an example case, consider how metabolic fluxes resulting from growth

on xylose compare with glucose-growth fluxes (Figure 4.5A). Glucose import and conversion reactions have lower flux, while xylose import/conversion reactions have higher flux, and these differences impact the metabolic flux network as a whole, e.g. via increased non-oxidative pentose phosphate pathway (PPP) activity from xylose's conversion to xylulose 5-phosphate. Expression profiling could be used to perform such analyses or support modeling-derived results (e.g., RNAseq data show that the glucose PTS permease *ptsG* decreased 2.4-fold in expression upon xylose growth, while the xylulokinase *xylB* increased expression 120 fold), but scalar gene expression values do not give insight into reaction directionality, which can cause important network features to be overlooked. Expression of *pgi*, for example, changed by less than 2% in glucose vs. xylose growth, but modeling showed that, although flux magnitude didn't appreciably change, the direction did – increased non-oxidative PPP activity from xylose growth leads to increased fructose 6-phosphate levels, which drive flux "backwards" through *pgi* to glucose 6-phosphate in the direction of gluconeogenesis. A two-sample Kolmogorov-Smirnov test was applied to the modeling-determined fluxes to quantify a metabolic "distance" between glucose and each of the substrates, and this scaled as expected – glycerol and xylose are more similar to glucose than acetate is (Figure 4.5B).

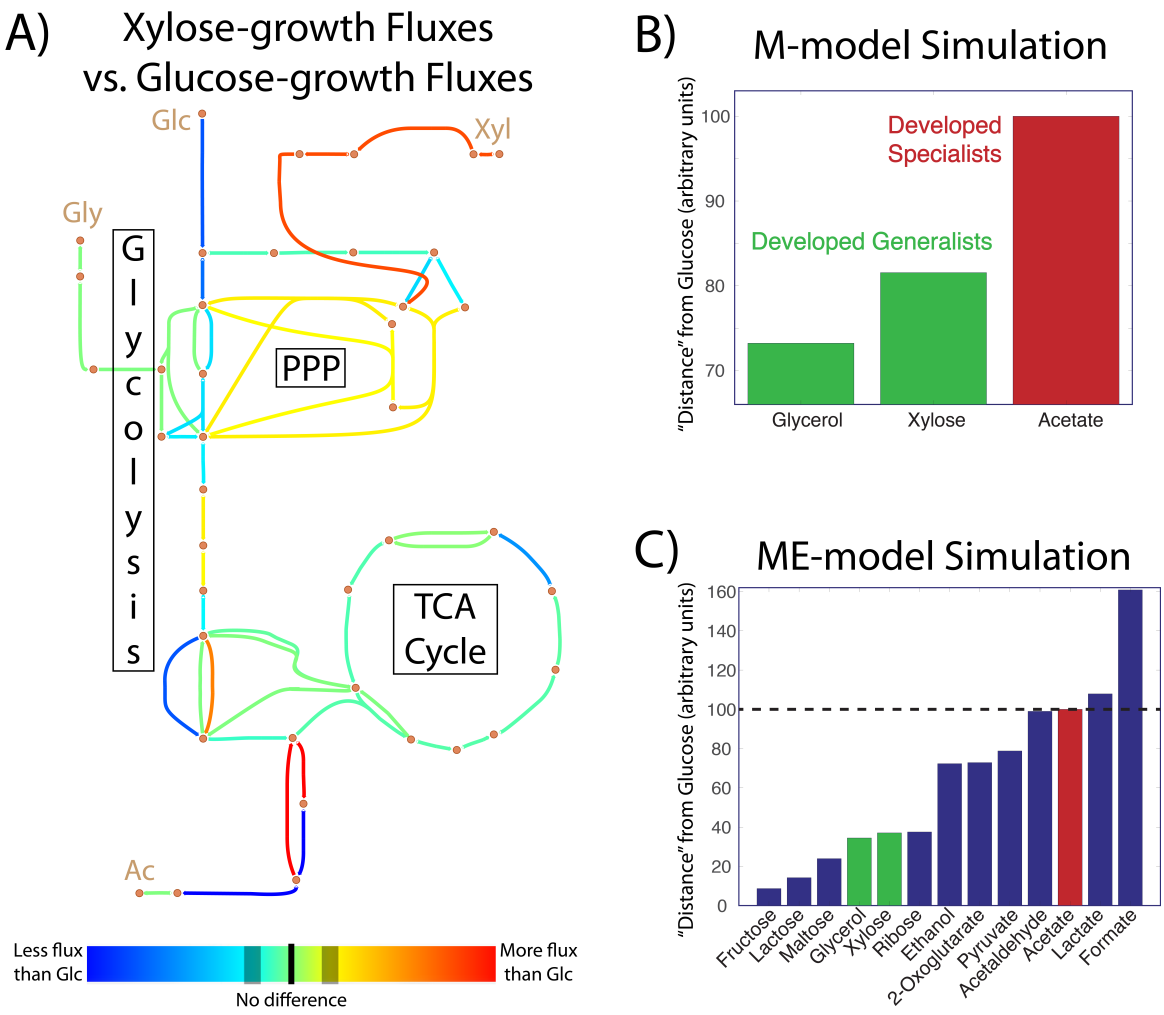



**Figure 4.5: Metabolic flux analysis of substrate differences** (A) Flux map of central carbon metabolism, with colors indicating the extent of difference in reaction fluxes between optimal growth on, as an example, glucose and xylose. (B) Substrate "distance" from glucose based on Monte Carlo sampling of the M-Model network, plotted against the observed frequency of specialist subpopulations arising in ALE strains under glucose/[substrate] switching. (C) Same as (B), but using Euclidean distance between ME-Model-predicted expression levels as the distance metric.

M-Models, although powerful tools for predicting and analyzing physiology, do not quantitatively predict gene expression, which can in certain circumstances lead to inaccurate predictions [118]. Monte Carlo sampling is one way to skirt this issue, but genome-scale models of metabolism that factor in gene expression and its concomitant energy costs have recently been developed, dubbed ME-Models [119]. This additional

model content allows for quantitative predictions on the optimal gene expression and flux state in a given environment without the need for random sampling. Taking the Euclidean distance between the expression state on glucose and on the other substrates is another way to quantify their extent of dissimilarity, and the results are in excellent agreement with M-model sampling (Figure 4.5C). Moreover, ME-models do not require physiological data on substrate uptake rates, so many compounds can be computationally tested. Dynamic ALEs on compounds with a lower glucose "distance" than glycerol (from these simulations, mono- and disaccharides) would likely lead to generalists, and compounds more distant than acetate (other carboxylic acids) would likely lead to specialists, but it is as of yet unclear at what intermediate substrate distance a dynamic evolution would begin to favor specialists over generalists. This study thus helps to establish the explanatory and predictive power of metabolic modeling for understanding why and under what circumstances generalists or specialists arise, but future experimental work will be needed to determine the extent and limitations of this efficacy.

## 4.3 Discussion

In this study, *E. coli* cultures were evolved for upwards of 1000 generations under environmental conditions in which the available carbon source alternated repeatedly between glucose and either glycerol, xylose, or acetate. These dynamically-evolved cultures reached fitnesses comparable to those of statically-evolved controls only ever grown on a single compound. Genetic analysis revealed that glucose/glycerol and glucose/xylose switchers adopted a generalist strategy, while glucose/acetate switching resulted in the development of specialist subpopulations for each of the two carbon

sources. Mutational comparison between static and dynamic strains highlighted genes important for robust growth on the various substrates, as well as ones targeted differentially depending on the static or dynamic nature of the evolutionary environment. The diauxic phenotype of dynamically-evolved strains varied across the substrates, but in one case lag phase was completely abolished whereas in statically-evolved controls it was not. Transcriptional analysis further highlighted the divergence in substrate-optimal expression states and resultant evolved expression shifts, and genome-scale metabolic modeling provided insight into the metabolic basis underlying substrate differences and generalist vs. specialist development.

Several important conclusions can be drawn from this study. Although the genetic basis for fitness improvement between static and dynamic conditions was mostly similar, several genetic regions stood out as being differentially targeted based on the temporal nature of the evolution environment. This is most noticeable in the *cspC/yobF* region, which acquired 5 unique mutations across the 6 total replicates of glc/xyl and glc/ac strains but only 1 mutation in a single acetate static strain. That such mutations occur disproportionately in dynamic conditions, and are not constrained to a single substrate pair, implies that this region influences substrate switching in general, perhaps through CspC's interactions with RNA polymerase complexes in response to stress [111]. Further analysis could establish the underlying mechanism for this influence, and find ways to leverage this knowledge for genetic engineering to produce strains with robust growth under dynamic conditions. However, simply looking at genes that mutate repeatedly does not provide the whole picture. Another striking finding from this study is the evolved diauxic behavior of a glucose/glycerol generalist, which eliminated its lag phase where

glucose and glycerol statically-evolved strains did not. In both the generalist and the glycerol-evolved strain the key mutations (Table 4.1) were in *glpK* and *rpoC*, genes known to be targeted under glycerol evolution [7]. Despite this genetic similarity, the different evolutionary histories selected for different mutations within the same genes, with a resultant phenotypic difference that would not be deduced from genotype alone. Additionally, diauxic lag phase elimination or duration reduction can significantly increase bioprocess fermentation efficiency [120]. The successful improvement in diauxic phenotype from both dynamic (e.g., glc/gly generalist, Figure 4.2B) and static (e.g., xylose-evolved strain, Figure 4.2C) ALE environments highlights the importance of utilizing both methods. ALE studies such as this one can thus help to expand the genetic knowledge base and indicate promising directions for genetic engineering of a desired phenotype, as well as naturally generating strains with industrially valuable phenotypes.

In addition to phenotypic and genetic analysis, transcriptomics and metabolic modeling helped to explain the observed evolutionary outcomes. Hierarchical clustering of static strain transcriptomes established the ordinality of substrate similarity (glucose → glycerol → xylose → acetate) that had been hinted at by the genetic and phenotypic ALE results. Generalist strains were found to shift their expression from the wild-type starting state closer to, but still distinct from, statically-evolved strains, while specialists adopted a different transcriptional strategy that moved them further from static strains than the wild-type. Where the generalists evolved to minimize the number and magnitude of gene expression shifts across two substrates, the specialists were not subject to this adaptive constraint given their tactic of coexistence and niche partitioning. Genome-scale modeling was also performed to examine the metabolic differences between the relevant

substrates, and perfectly recapitulated the experimentally determined substrate similarity hierarchy. These results indicate promising avenues of investigation for future studies. Firstly, evolving to static conditions can yield insight into the results of dynamic evolutions. The amount to which expression states differed among statically-evolved strains explained the development of subpopulations under dynamic environments, and as ALE studies increase in number there are more and more available data researchers can reference when designing studies of their own [5, 74]. However, when such data are lacking or it is prudent to avoid the time and resource costs of performing a static ALE, genome-scale metabolic modeling serves as a way to make these predictions without requiring starting data. The propensity for specialist subpopulation development can be deduced from the modeling-quantified substrate "distances," and ALE studies can be designed accordingly depending on the desired outcome. If a single dominant genotype were desired, such as when optimizing a genetically engineered strain [121], an environment favoring generalist development could be selected; if overall culture performance were instead the important factor, then an environment favoring specialists would not need to be avoided, allowing naturally evolved specialists to substitute for artificially engineered microbial consortia that would have the same collective phenotype [122, 123].

Overall, the data presented herein provide insight into adaptive strategies and evolutionary outcomes in dynamic environments, and demonstrate the efficacy of various data types for analyzing or designing such studies. Dynamic environments present a much more complicated selection pressure than static alternatives, with increased environmental heterogeneity known to lead to greater population heterogeneity [107]. As

dynamic ALE studies increase in number [98, 101] it is essential that appropriate experimental tools are in place to properly guide analyses and assess outcomes. Moreover, the dynamically-driven development of generalists or specialists is of clinical importance in regards to antibiotic resistance and treatment regimes [124]. The genome-scale models used in this study to quantify metabolic variation due to growth on different substrates can also model the network perturbations caused by antibiotics targeting specific biochemical reactions [125]. With modeling-driven predictions and -omics data follow-up characterizations after ALE experiments, strides can be made in both basic evolutionary research and applied clinical and biotechnological studies.

## 4.4 Materials and Methods

### 4.4.1 Adaptive Laboratory Evolution and Phenotypic Profiling

Strains were evolved in an automated system that tracked growth rates and propagated cultures in constant exponential growth phase, as described previously [126]. Starting with wild-type *Escherichia coli* K-12 MG1655 (ATCC 4706), cultures were serially propagated (100 μL passage volume) in 15 mL (working volume) tubes of M9 minimal medium kept at 37°C and well-mixed for full aeration. Cultures were propagated once reaching an Optical Density at 600 nm of 0.3 (Tecan Sunrise plate reader; equivalent to an $OD_{600}$ of ~1 on a typical instrument with 1 cm path length), a point at which nutrients were still in excess and exponential growth was still occurring (confirmed with growth curves and HPLC measurements). The M9 medium contained

either 4 g/L glucose, 4 g/L xylose, 4 g/L acetate, or 0.2% (by volume) glycerol. Dynamic cultures were alternately passed between glucose medium and one of the three alternative media types, while static cultures were only ever grown on a single type of medium. For dynamic cultures, average growth period in glucose tubes (time from inoculation to passage) decreased from 8.2 hours to 5.3 hours over the course of the ALE experiment, while average xylose growth period decreased from 9.1 to 6.0 hours, glycerol decreased from 10.4 to 5.8 hours, and acetate decreased from 22.5 to 12.9 hours. Lag phases were not evident when passing cultures between alternating substrates, likely because lag would occur immediately after passage and before $OD_{600}$ values were detectable. Diauxic growth tests were performed under identical conditions to the ALE experiment, but M9 medium was used containing 5 mM glucose + 5 mM xylose or acetate or glycerol.

### 4.4.2 DNA and RNA sequencing

Genomic DNA, either clonal or population, was isolated using the Machery-Nagel Nucleospin Tissue Kit, following the manufacturer's protocol for use with bacterial cells. The quality of isolated genomic DNA was assessed using Nanodrop UV absorbance ratios. DNA was quantified using the Qubit dsDNA high-sensitivity assay. Paired-end whole genome DNA sequencing libraries were generated using Illumina's Kappa Kit and run on an Illumina Miseq platform with a PE600v3 kit. The generated DNA sequencing fastq files were processed with the breseq computational pipeline [82] and aligned to the *E. coli* genome (NCBI accession NC_000913.3) to identify mutations. For population sequencing, evolved endpoint populations were used to inoculate a tube of medium with the desired carbon source and DNA was then harvested following growth enrichment

overnight. Mean read depth for each population was at least 110x, and only mutations with greater than 15% population frequency in at least one growth condition were examined. Mutations differing by more than 1.5x in frequency across different carbon source growth enrichments were taken as evidence of subpopulations (Supplementary File 2). DNA sequencing data from this study are available from the Sequence Read Archive database (SRP103966).

RNA sequencing data were generated under conditions of aerobic, exponential phase growth on M9 minimal medium plus the relevant carbon substrate at the concentrations used in the ALE experiment. Cells were harvested using the Qiagen RNA-protect bacteria reagent according to the manufacturer's specifications. Prior to RNA extraction, pelleted cells were stored at -80°C, then thawed and incubated with lysozyme, SuperaseIn, protease K, and 20% sodium dodecyl sulfate for 20 min at 37°C. Total RNA was isolated and purified using Qiagen's RNeasy minikit column according to the manufacturer's specifications. Ribosomal RNA (rRNA) was removed using Epicentre's Ribo-Zero rRNA removal kit for Gram-negative bacteria. A KAPA Stranded RNA-Seq Kit was used to generate paired-end, strand-specific RNA sequencing libraries, which were then run on an Illumina HiSeq. RNAseq reads were aligned to the *E. coli* genome using bowtie2 [65] and FPKM values were calculated with cufflinks [67]. Each sample had at least 80x mean read depth coverage. Sample normalization was performed with cuffnorm and differential expression levels were quantified via cuffdiff [127]. RNA sequencing data from this study are available from the Gene Expression Omnibus database (GSE97944).

### 4.4.3 *In silico* modeling

Monte Carlo sampling of M-model flux distributions under different substrate growth conditions was performed on the most current genome-scale model of *E. coli* metabolism, *i*JO1366 [128], using the Matlab COBRA Toolbox [129], as described previously [117]. Substrate uptake rates for the different carbon compounds were set to wild-type values (Supplementary File 3) and the allowable growth rate to within 10% of model-determined optimum, and Monte Carlo sampling was performed with the sampleCbModel function using default parameters. A two-sample Kolmogorov-Smirnov test statistic was used to pairwise compare, for every reaction in the model, the difference in flux distribution between xylose- or glycerol- or acetate-growth and glucose-growth. The cumulative sum of test statistics for every reaction led to a quantitative value of metabolic "distance" from glucose, and values were normalized such that acetate's distance was 100. Flux differences were visualized by mapping model outputs to a metabolic pathway map (Figure 4.5A) via the Escher tool [130].

A genome-scale model of *E. coli* metabolism *and* gene expression (ME-Model), *i*LE1678-ME [131], was used to simulate growth on each of the four carbon substrates used in this study, as well as several other common growth substrates. To model each condition, the uptake of the growth substrate was unconstrained and other carbon substrates' uptakes were set to 0. The remaining default ME-model parameters are optimized to model the growth of a laboratory evolved *E. coli* K-12 MG1655 strain, so they remained set to their default values. The ME-model was simulated by computing the maximum feasible growth rate of the model under the imposed *in silico* conditions via a bisection procedure [132] that uses a quadruple-precision version of the MINOS

optimizer [133]. A single ME-model simulation provides a prediction of the transcriptome, proteome, and metabolic flux state required for the cell to grow optimally. Using these values, the metabolic "distance" of a compound from glucose was quantified by calculating the Euclidean distance between ME-predicted translation reaction fluxes (proteome), and values were normalized such that acetate's distance was 100.

Chapter 4 is a reprint of a published manuscript: Sandberg TE, Lloyd CJ, Palsson BO, Feist AM. 2017. "Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies." *Applied and Environmental Microbiology*, 83(13):e00410-17. The dissertation author was the primary author of the paper and was responsible for the research

# Chapter 5

# Synthetic Cross-Domain Gene Replacement and Evolutionary Assimilation of Major Glycolytic Enzymes into *E. coli*

## 5.1 Introduction

The ability of life forms to incorporate and use foreign DNA is an incredibly important process, underlying much of the phenotypic flexibility of Earth's biome. In natural settings this biological feature enables species diversification via Horizontal Gene Transfer (HGT), and in applied settings it can be harnessed to engineer organisms for a desired purpose, such as heterologous protein expression.

For an HGT event to be selected for in the wild the new DNA sequence must provide a fitness advantage via which HGT strains can outcompete non-HGT strains in a particular niche, after which adaptive mutations further refine the functionality of the HGT product in its new host. This process underlies much of the rise and spread of antibiotic resistance [134]. Unfortunately, while facilitating HGT in the wild, evolution tends to work against the successful use of heterologous expression for biotech purposes.

This is due to the vastly different role HGT plays in biotech vs. nature – human genetic engineering of organisms is generally done to induce a desirable phenotype, with concomitant fitness costs that render the engineered cells easily outcompeted by mutants that purge the foreign DNA. Moreover, simply providing the genetic instructions for proteins does not mean they will be produced in proper amounts, with proper functionality, or without negatively impacting cell physiology. Techniques have been developed to lessen these issues, such as codon optimization to increase expression in the new host or pathway engineering to force use of the foreign DNA [135], but biological parts remain much less 'plug and play' than would be ideal [136].

Given the importance of understanding principles governing cross-species genetic interchangeability, several studies have investigated this matter. Kachroo *et al.* performed large scale essential gene replacements in yeast using both human [137] and bacterial [138] donor DNA, finding pathway-specific biases in outcome and swap successes for roughly half of all genes targeted. Functional gene replacement is an important dataset, but pairing with evolution allows for greater insight into the physiological consequences of the foreign DNA. Lind *et al.* replaced ribosomal genes in *S. typhimurium* with microbial orthologs, and within a few hundred generations of laboratory evolution found gene amplifications aimed at increasing orthogene copy number as a way to ameliorate fitness defects [139]. Kacar *et al.* replaced the essential elongation factor *tufB* in *E. coli* with an ancestral variant, and evolution similarly selected for upregulation [140]. Bershtein *et al.* replaced the essential *folA* gene in E. coli with orthologs from 35 close bacterial relatives and evolved, finding that fitness defects were frequently compensated by protease-deactivating mutations that increased intracellular orthogene levels [141].

The aforementioned studies begin to shine light on the factors governing gene replaceability, but the influence of particular variables remains difficult or impossible to deconvolute from existing data. Codon optimization was typically performed on the foreign genes before insertion, a critical change that limits applicability of observed results to natural HGT. Essential genes were also the main targets for replacement, precluding any outcome in which an initially non-functional swap could evolve functionality, and vastly limiting the pool of organisms from which an orthogene could be taken without inducing lethality in the new host. Additionally, the *in vivo* role for many of the chosen genes involves protein-protein interactions and complex formation, adding confounding factors that muddy mutational interpretation.

To answer remaining unsolved questions of how a cell differentiates "self" from "non-self" components, an ALE experiment was carefully designed and implemented. Eight unique gene-swapped strains of *E. coli* were constructed, evolved for hundreds to thousands of generations, and analyzed to discover principles governing success or failure of foreign gene assimilation.

## 5.2 Results

### 5.2.1 Strain Design and Construction

A number of factors constrained the experimental design in terms of gene targets, donor organisms, and method of strain construction. The genes *pgi* and *tpiA* were ideal targets for a variety of reasons – neither gene is essential, but KO causes a fitness decrease of ~70%, enabling viability of strains with initially non-functional swaps but providing extremely strong selective pressure to get the swapped gene working. Both

genes are well studied, with known crystal structures, enzymatic mechanisms, physiological consequences of the KO, and even adaptive outcomes from evolution of KO strains [11, 142-144]. Neither gene has significant protein-protein interactions or necessary cofactors, reducing factors that could influence mutational results and complicate analysis. To further minimize potential confounding factors, strain construction involved scarless chromosomal replacement, from start codon to stop codon, with the coding sequence of the foreign ortholog not subjected to any codon optimization. This method of construction preserves native sequence context, leaving the promoter, 5'-UTR, terminator, and genomic location completely unchanged.

Orthogene donor organisms were selected so as to span a range of dissimilarities from *E. coli*: the fellow gammaproteobacterium *Vibrio cholerae* (Vch), the alphaproteobacterium *Brucella melitensis* (Bme), the hyperthermophilic archaeum *Pyrobaculum aerophilum* (Pae), and the mammalian eukaryote *Homo sapiens* (Hsa) (Table 5.1, Figure 5.1B). This spectrum of swaps allows us to distinguish between adaptive outcomes and mutations specific to a particular gene, specific to a particular species' DNA, or general aids to heterologous expression. Spanning all three domains of life, the *pgi* and *tpiA* orthologs from these organisms have notably different GC content (Figure 5.1C), codon usage (Figure 5.1D), and amino acid composition (Figure 5.1E) when compared to the native *E. coli* genes.

**Table 5.1: Donor species and gene properties**

| | | | | Gene-specific Properties | | | |
|---|---|---|---|---|---|---|---|
| **Organism Properties** | | | | ***pgi*** | | ***tpiA*** | |
| **Species** | **Domain** | **Type** | **Optimal Growth Temperature** | **Protein Alignment Score** | **Codon Adaptation Index** | **Protein Alignment Score** | **Codon Adaptation Index** |
| *Escherichia coli* | Bacteria | γ-proteobacteria | 37 °C | 2886 | 0.79 | 1283 | 0.77 |
| *Vibrio cholerae* | Bacteria | γ-proteobacteria | 37 °C | 2280 | 0.74 | 850 | 0.73 |
| *Brucella melitensis* | Bacteria | α-proteobacteria | 37 °C | 1366 | 0.71 | 448 | 0.66 |
| *Homo sapiens* | Eukarya | mammalian | 37 °C | 1864 | 0.62 | 487 | 0.59 |
| *Pyrobaculum aerophilum* | Archaea | crenarchaeota | 100 °C | 40 | 0.53 | 161 | 0.54 |

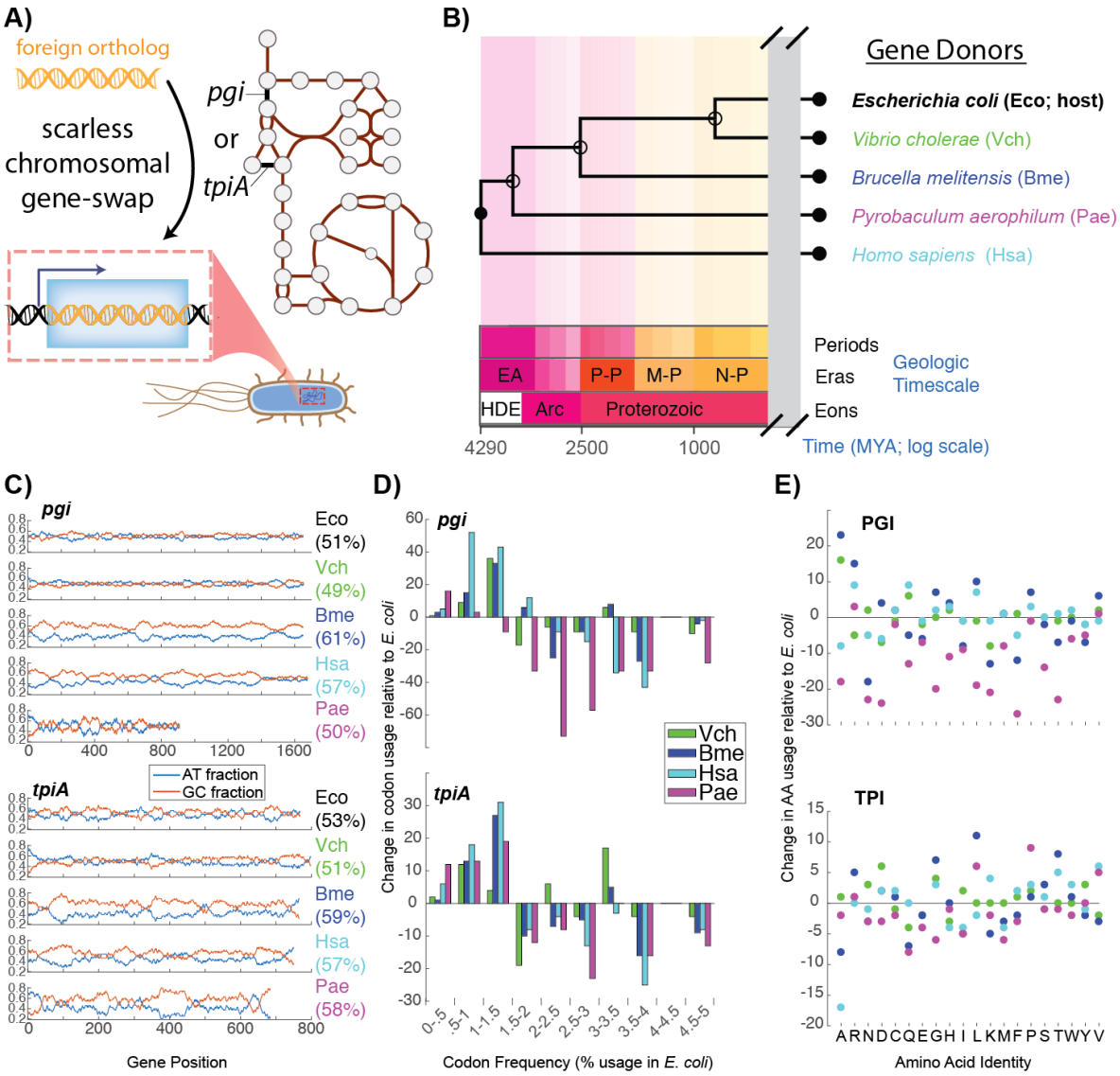



**Figure 5.1: Gene-swap strain construction and properties** (A) Schematic of metabolic network and method of strain construction. (B) Phylogenetic distance between *E. coli* host strain and gene donor species. (C) GC content of native and donor gene sequences (GC% total in parentheses) (D) Histogram of the change in codon usage resulting from replacement of native sequences with foreign versions (E) Change in protein's amino acid usage resulting from replacement of native sequences with foreign versions

## 5.2.2 Pre- and Post-Evolution Physiology

Initial physiology of gene-swapped strains was a function of the donor species - both Vch swaps had wild-type growth rates, both swaps for Hsa and Pae had KO growth

rates, and both swaps for Bme had rates intermediate to KO and wild-type. However, the BmePgi swap grew only 20% faster than the KO (60% slower than WT), while the BmeTpi swap grew 185% faster (17% slower than WT), demonstrating significant gene-specific differences in interchangeability even from the same donor organism. Given the stochastic nature of evolution and uncertain outcomes for these atypical strains, a large number of independent replicates was desirable – more than 65 independent lineages were evolved to optimize exponential-phase growth on glucose minimal media for ~30 days (Figure 5.2A). This experimental length was chosen so that fitness trajectories would be dominated by sequential selective sweeps of large-effect beneficial mutations [145]. Other studies serve as crucial comparison points – under identical selection conditions, adaptive outcomes for wild-type [5, 126], Δ*pgi* [11, 142], and Δ*tpiA* [144] *E. coli* strains are well established, and additional controls run alongside the gene-swaps recapitulated these results (Figure 5.2B).

Based on fitness trajectories from the ALE experiments and final growth rates reached by the populations, each independent lineage was classified as either a 'success' or 'failure' in terms of foreign gene assimilation (Figure 5.2C). Such classification is greatly facilitated by the intrinsic growth rate gap between *pgi* or *tpiA* deficient strains and the wild-type, even when comparing evolved KOs to the unevolved wild-type. ALE outcomes were sometimes, but not always, consistent across replicate lineages founded from the same ancestor – while 10/10 BmePgi swaps and 0/10 PaePgi swaps were successful, HsaPgi swaps were 60% successful and 40% failures (Figure 5.2D). Notably, a single HsaTpi lineage was the only failure across all *tpiA* swaps, establishing significant gene-specific differences in ease of orthogene assimilation.

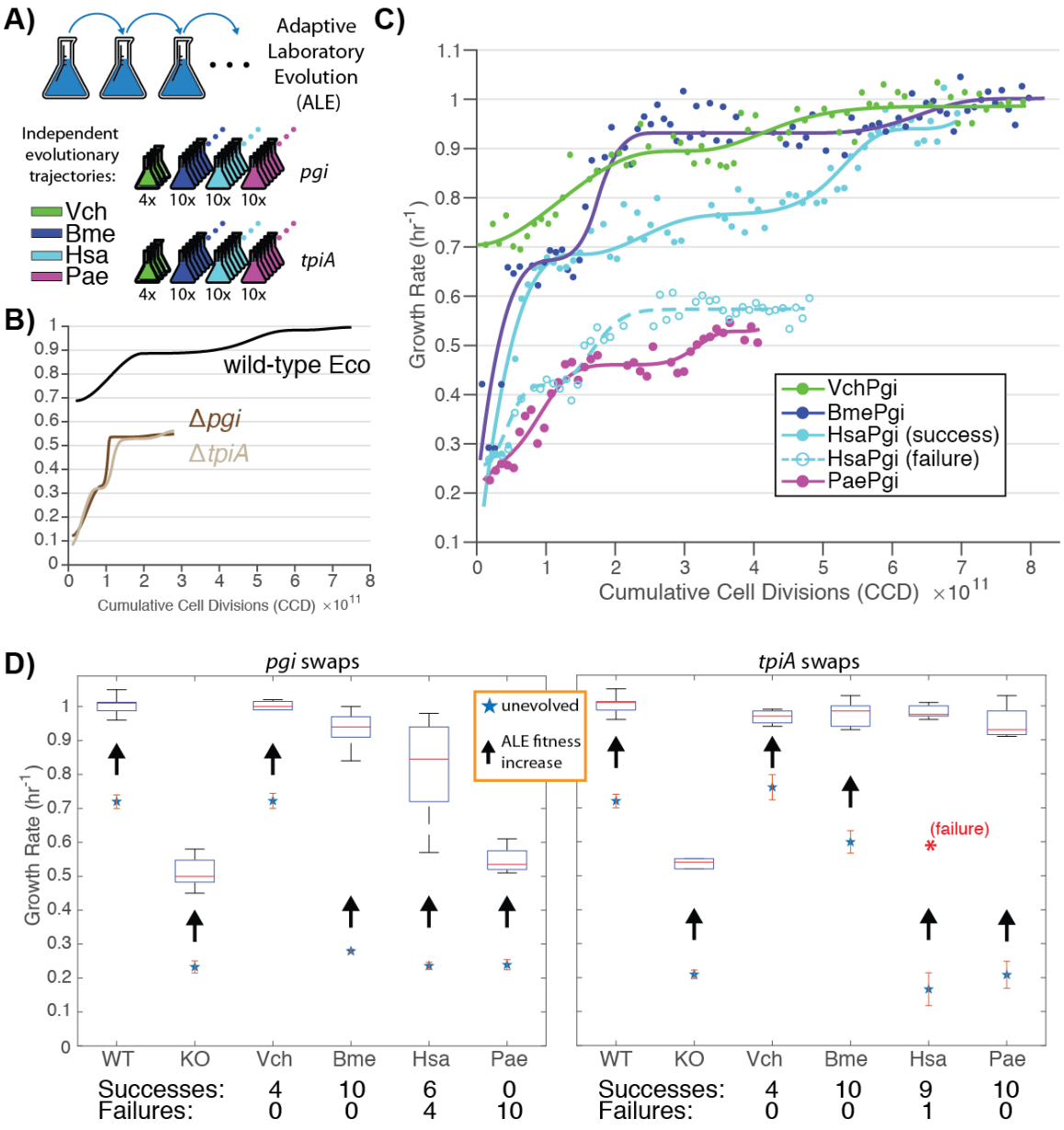


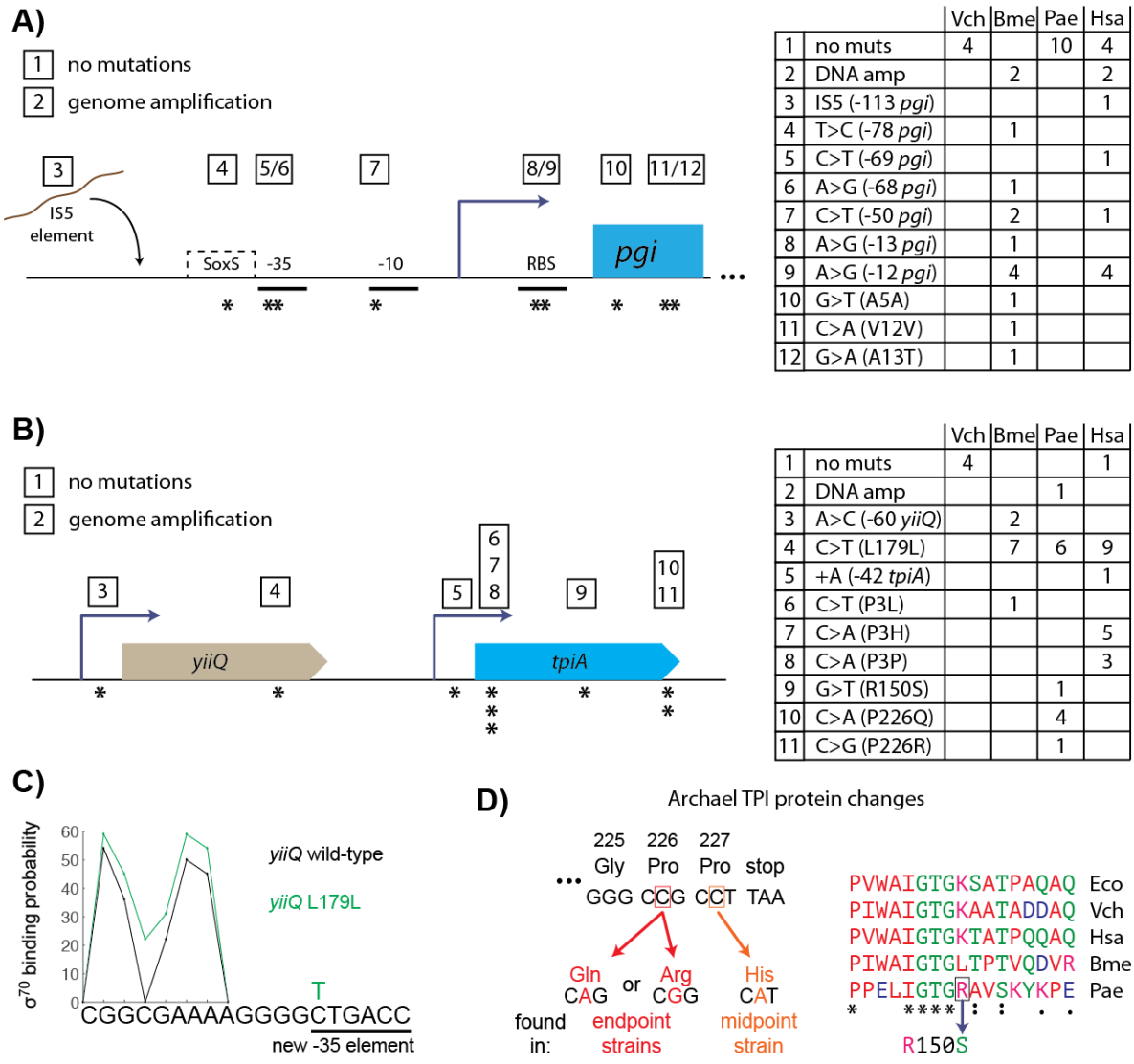

**Figure 5.2: Fitness trajectories and evolutionary outcomes** (A) More than 65 independent lineages were serially propagated in glucose minimal media for ~30 days (B) Typical fitness trajectories for wild-type, Δ*pgi*, and Δ*tpiA E. coli* strains evolved on glucose minimal media (C) Individual gene-swap lineages could be characterized as a 'success' or 'failure' of orthogene assimilation based on final growth rate reached (D) Starting growth rates of gene-swapped strains and box plots of ALE growth rate outcomes across the independent replicates

## 5.2.3 Mechanisms of Orthogene Assimilation

Evolved endpoint clones were isolated and whole genome sequenced to determine mechanisms of adaptation. Strikingly, but not surprisingly, every single "successful' replicate (reaching a growth rate above 0.75 /hr) had one or more mutations in or around the foreign gene, barring *Vibrio cholerae* swaps which didn't require any mutations to enable functionality in *E. coli* (Figure 5.3A-B). Genome amplifications, an established adaptive mechanism for orthogene upregulation [139], were one of the least prevalent mutational types for both *pgi* and *tpiA*. In the case of *pgi*, far more common were mutations targeting the ribosome binding site or promoter. The exact same A→G SNP 12 basepairs upstream of the start codon occurred independently eight separate times across both Bme and Hsa swaps, changing the RBS from AGAAGA to AGAGGA, closer to the canonical Shine-Delgarno sequence. After RBS mutations, promoter mutations were the most frequent method of *pgi* upregulation, with SNPs falling exactly in the -10 or -35 sites, again with independent, identical mutations across strains. Evolved *tpiA* swap strains had markedly different adaptive mutations compared to *pgi* swaps – rather than promoter or RBS changes, by far the most common mutation was a silent SNP in *yiiQ*, the gene upstream of *tpiA*, occurring independently more than 20 times. Overall, changes to the foreign coding sequence were less common than proximal mutations, and fell into two main types – missense SNPs in the archaeal PaeTpi, and a number of N-terminal, mostly synonymous SNPs across a variety of swapped strains.

Existing literature and gene annotation allows for easy interpretation of RBS and promoter mutations, but other mutations require more detailed analysis. For the *yiiQ* L179L SNP to aid in orthogene assimilation it must increase levels of the foreign *tpiA*, and it appears to accomplish this by creating a new promoter binding site. Analysis with

promoter prediction tools [146] reveals the increased chance for RNA polymerase binding which the C→T mutation creates (Figure 5.3C). This mechanism has been documented in a study before [147], and highlights the importance of genome location and local context on adaptive outcomes. As far as coding sequence changes, the PaeTpi swaps repeatedly acquired SNPs to the penultimate amino acid. Given that the protein ends with two prolines, this is most consistent with a mechanism of increasing expression by reducing poly-proline ribosomal stalling [148], a conclusion supported by the isolation of growth-improved clones stemming from changes to either of the two proline residues (Figure 5.3D). The archael tpi swap is also characterized by the only observed mutation that likely alters enzyme activity rather than expression level – residue 150 is immediately adjacent to the conserved proton acceptor region of the active site.




| | | Vch | Bme | Pae | Hsa |
|---|---|---|---|---|---|
| 1 | no muts | 4 | | 10 | 4 |
| 2 | DNA amp | | 2 | | 2 |
| 3 | IS5 (-113 *pgi*) | | | | 1 |
| 4 | T>C (-78 *pgi*) | | 1 | | |
| 5 | C>T (-69 *pgi*) | | | | 1 |
| 6 | A>G (-68 *pgi*) | | 1 | | |
| 7 | C>T (-50 *pgi*) | | 2 | | 1 |
| 8 | A>G (-13 *pgi*) | | 1 | | |
| 9 | A>G (-12 *pgi*) | | 4 | | 4 |
| 10 | G>T (A5A) | | 1 | | |
| 11 | C>A (V12V) | | 1 | | |
| 12 | G>A (A13T) | | 1 | | |

| | | Vch | Bme | Pae | Hsa |
|---|---|---|---|---|---|
| 1 | no muts | 4 | | | 1 |
| 2 | DNA amp | | | 1 | |
| 3 | A>C (-60 *yiiQ*) | | 2 | | |
| 4 | C>T (L179L) | | 7 | 6 | 9 |
| 5 | +A (-42 *tpiA*) | | | | 1 |
| 6 | C>T (P3L) | | 1 | | |
| 7 | C>A (P3H) | | | | 5 |
| 8 | C>A (P3P) | | | | 3 |
| 9 | G>T (R150S) | | | 1 | |
| 10 | C>A (P226Q) | | | 4 | |
| 11 | C>G (P226R) | | | 1 | |

**Figure 5.3: Mutational mechanisms of orthogene assimilation** (A) *pgi*-proximal mutations found in gene-swapped endpoint ALE strains (B) *tpiA*-proximal mutations found in gene-swapped endpoint ALE strains (C) New promoter creation via silent SNP upstream of *tpiA* (D) Orthogene coding sequence changes found in evolved PaeTpi strains; C-terminus of the gene and multiple sequence alignment of the active site region

## 5.2.4 mRNA Stem-loop SNPs as an Adaptive Mechanism for Tuning Expression

Other than PaeTpi nonsense SNPs, the only observed coding sequence changes were a number of N-terminal SNPs across multiple different strains, which often didn't alter the amino acid sequence. The ability of synonymous mutations to nevertheless increase fitness has been observed before, but studies examining such occurrences have been unable to deduce the exact causal mechanism – tuning codon frequency, altering internal ribosome sites, or changing mRNA secondary structure all remain potential explanations [149, 150]. Fortunately, the large sample size of this study, coupled with computational analysis tools [151], allows us to confidently deduce the adaptive mechanism of these mutations – they appear to target stem-loops, or hairpins, in the mRNA secondary structure, destabilizing them and thereby "opening up" the transcript for increased ribosomal readthrough. In the case of the human *tpiA* swap, the foreign coding sequence binds to the native 5'-UTR and the P3P and P3H SNPs (and an S4S SNP found in a midpoint strain) target this stem-loop for destabilization (Figure 5.4A). This also explains the observed +A mutation 42 bp upstream of the stop codon – it similarly destabilizes the stem-loop, but on the opposite side as the coding SNPs and via increased stabilization of a stem-loop-adjacent, unstructured region. SNPs found in evolved BmePgi strains are even more strikingly aimed at stem-loop destabilization, this time due to binding of the coding sequence with itself rather than the native 5'-UTR. A5A and V12V target the exact same rung of the primary stem-loop, but on opposite sides, and A13T targets an adjacent rung (Figure 5.4B-C)

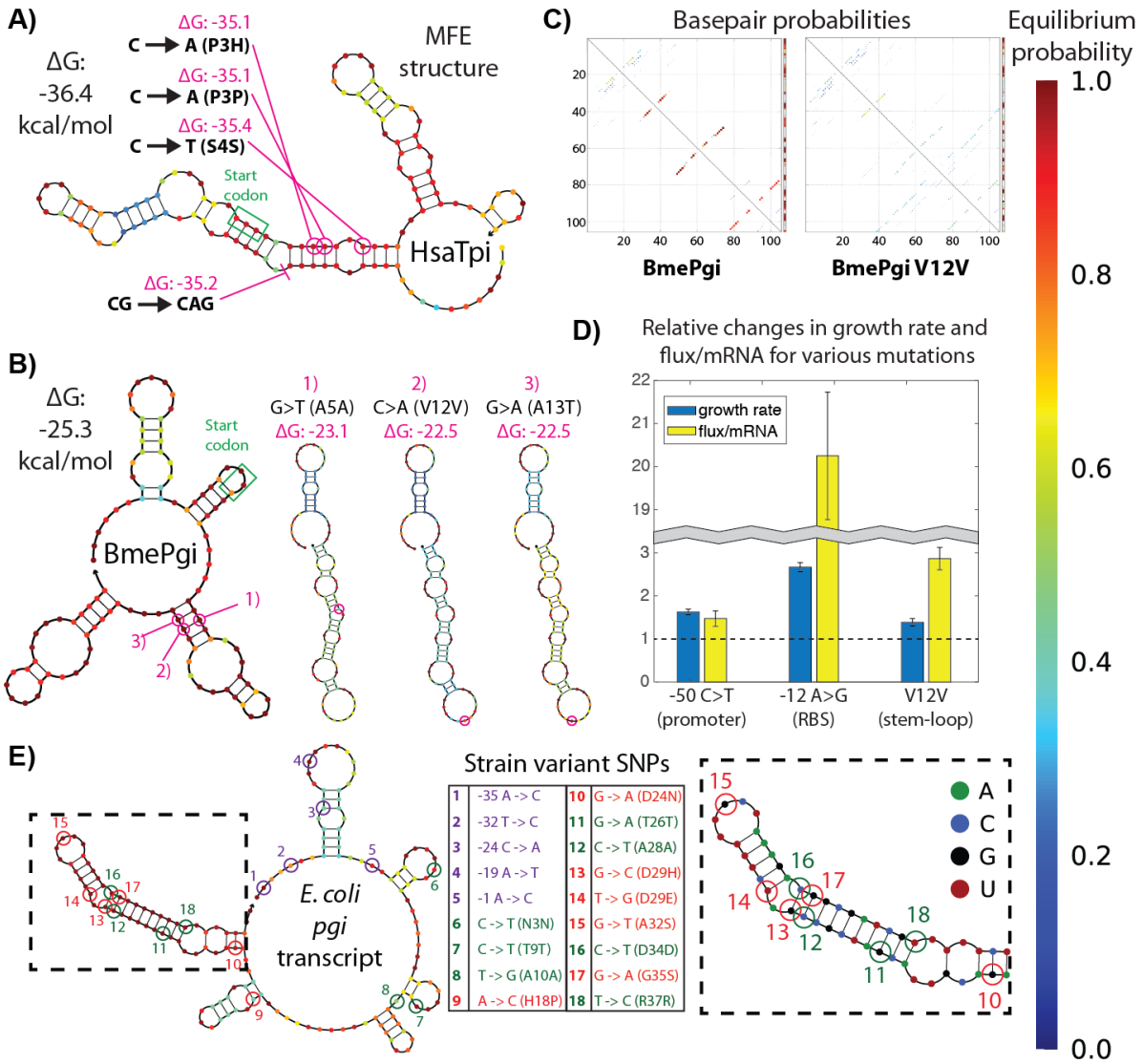



**Figure 5.4: SNPs targeting mRNA stem-loops** (A) Minimum free energy (MFE) structure at 37 °C of *tpiA* transcript for HsaTpi swap, with observed mutations (B) MFE structure at 37 °C of *tpiA* transcript for BmePgi swap, with observed mutations and altered MFEs (C) Equilibrium pair probabilities at 37 °C for unaltered BmePgi and BmePgi V12V transcripts; the less red the coloration, the more unstructured the MFE (D) Mutation-induced changes in growth rate and flux/mRNA for different mutational types (E) SNP accumulation in *pgi* across 924 sequenced *E. coli* strain variants

To empirically validate this deduced mechanism of phenotypic influence, we performed pairwise characterizations of *pgi* expression level and enzymatic activity in strains genetically identical except for single mutations of interest, thus enabling causal

establishment. Promoter SNPs should increase both mRNA expression levels and flux, while RBS and stem-loop SNPs should increase protein expression, and thus flux, but by virtue of falling within the transcript have minimal impact on mRNA expression. This expectation is born out by the results – the repeatedly observed -10 element promoter SNP increases growth rate via roughly equal fold changes in flux and mRNA, while the RBS SNP significantly increases flux/mRNA, and the V12V SNP increases it to a smaller extent (Figure 5.4D).

We next sought to establish if these stem-loop SNPs are potentially a widespread mutational mechanism, or merely a consequence of the atypical transcripts caused by the gene-swaps. In a survey of 924 *E. coli* strain variants with whole genome sequences available, SNPs within the first 40 codons of *pgi* were found to accumulate preferentially in the area of strongest mRNA secondary structure, specifically a 5-rung GC segment (and thus strongest region) of a longer stem-loop (Figure 5.4D). When taken together with another study examining 'nucleotide compatibility' across pathogenic *E. coli* strains, in which *pgi* stood out as the gene most frequently targeted by mutations [152], we begin to see how these stem-loop SNPs may in fact be underappreciated drivers of adaptation rather than neutral signatures of drift. Indeed, mRNA structural analysis of N-terminal synonymous mutations found in other ALE experiments gives further evidence for their phenotypic impact – e.g., *fpr* N14N for increased ferredoxin expression under ROS stress [153], or *carB* L11L for increased expression following KO of *carA* [154].

### 5.2.5 Adaptive Dynamics and Genome Rearrangements

To probe the evolutionary dynamics governing observed fitness trajectories and adaptive outcomes, we isolated and whole-genome sequenced multiple midpoint strains from every ALE lineage, in addition to the previously sequenced endpoints. With 200+ genetically distinct strains, in many cases we could decipher the full stepwise acquisition order for every mutation found in an endpoint, while in others clonal interference was observed as different strains alternated dominance in the evolving ALE populations.

Although mutations to the swapped gene were the ultimate determinate of 'success' or 'failure' of orthogene assimilation, a number of other genes exhibited strongly causal influence via three or more independent mutational hits across this or reference studies (Table 5.2). In many cases the mutational occurrences reinforced conclusions drawn from initial physiology and adaptive outcomes – for example, 3/4 VchPgi strains got *hns/tdk* IS element mutations, a causal mechanism only observed for wild-type glucose evolutions [5, 126], highlighting the negligible influence of the Vch swap. Moreover, 'failure' lineages, in addition to lacking orthogene mutations, had mutations characteristic of evolved KO strains – for example, the gene *sthA* didn't mutate in any 'successful' *pgi* swaps, but did in two of the failures (1 HsaPgi, 1 PaePgi) and multiple Δ*pgi* controls [11]. Although only a single HsaTpi swap was a failure, it likewise had unique adaptive signatures characteristic of Δ*tpiA* evolutions – *ptsG*, *galR*, and *nemR* all mutated [144].

**Table 5.2: Independently recurring mutations**

| | Causal influence in ALE of different genetic backgrounds? | | | Gene swap occurences | |
|---|---|---|---|---|---|
| **Gene** | **WT**[a] | **Δ*pgi***[b] | **Δ*tpiA***[c] | ***pgi*** | ***tpiA*** |
| *hns/tdk* | y | | | 3 | 1 |
| *pyrE/rph* | y | | | 16 | 19 |
| *rpoB* | y | y | y | 12 | 10 |
| *rpoC* | y | y | y | 14 | 9 |
| *rpoA* | | y | | 14 | 7 |
| *icd* | | y | | 7 | 4 |
| *crp* | | y | | 5 | 1 |
| *gnd* | | y | | 5 | 0 |
| *cyaA* | | y | | 4 | 0 |
| *hfq* | | y | | 4 | 0 |
| *sthA* | | y | | 2 | 0 |
| *sgrR* | | y | | 2 | 0 |
| *oxyR* | | | y | 0 | 7 |
| *pykF* | | | y | 1 | 16 |
| *nusA* | | | | 4 | 3 |
| *hemA/prfA* | | | | 1 | 2 |

[a] Refs: [5, 126], [b] Refs: [11, 142], [c] Ref:[144]

Of note are multiple mutations that cannot be uniquely classified as success- or failure-characteristic, but which are found across a variety of strains. For example, *icd*, *cyaA,* and *crp* are known to accumulate mutations in Δ*pgi* evolutions, and were unsurprisingly found mutated in multiple failure lineages, but notably also in several successful lineages. Similarly, although *tpiA* swaps were all successful but for one, *oxyR* and *pykF* were still found to mutate as observed in Δ*tpiA* ALEs. Strikingly, the exact same mutation was sometimes acquired – the *crp* A152E SNP was found in a failure PaePgi lineage, but also a successful BmePgi one. Such shared mutations across both successful and KO/failure lineages could represent escape events from local fitness maxima after an ultimately successful lineage initially began a KO-like adaptive walk, or

they could indicate a shared underlying physiological state resulting from either absence of, or deficit in, *pgi* or *tpiA* flux, towards which the mutations prove beneficial regardless of lineage success or failure. Our high frequency sampling of the fitness landscape revealed temporal mutation dynamics providing explicit instances of both. As examples, in one BmePgi lineage a *cyaA* mutation fixed after the strain had already started down a 'success' trajectory by acquiring two different upstream *pgi* SNPs. In a different HsaPgi lineage, the first mutation was a specific *rpoA* SNP characteristic of *pgi* KO/failures, but it was followed by the -12 *pgi* RBS SNP, and ultimately lineage 'success.'

To examine if successful orthogene assimilation was possible even in the presence of multiple KO-characteristic mutations, we performed 'continuation ALEs' starting from endpoint strains isolated from various failed lineages, evolving them for a further 30 days (Figure 5.5A) The additional evolutionary time enabled success for the HsaTpi strain that was the only *tpiA* swap failure, and for a midpoint BmePgi strain that started with two KO-characteristic mutations. However, continuation did not always guarantee success, as HsaPgi and PaePgi strains remained 'failures' despite minor fitness improvements.

Surprisingly, our extensive midpoint sequencing found that genome amplifications were much more common than endpoint sequencing alone revealed. Although only a single additional *pgi* amplification was found in a midpoint HsaPgi strain, complementing the amplifications found in two BmePgi and two HsaPgi endpoints (Figure 5.5B), *tpiA* amplifications were rampant – more than 90% of clones isolated across Bme/Hsa/PaeTpi lineages after their first fitness trajectory growth rate jump had amplifications. As the ALE experiment progressed these *tpiA*-amplified strains were

ultimately outcompeted by strains with more parsimonious methods of upregulation (*yiiQ* L179L and mRNA stem-loop SNPs), leaving only a single PaeTpi endpoint with a persistent amplification.

Examination of the type and location of genome amplifications helped to explain the swap-specific differences in evolutionary outcomes for orthogene copy number. The location of *pgi* on the chromosome leaves it open to homology-mediated genome amplification of various types – flanking repetitive extragenic palindromic, or REP, sequences facilitated several of the observed amplifications, as well as sequence homology between flanking genes (Figure 5.5B). Minimal homology was needed – as few as 10 identical bp between *metH* and *mdtP* could nonetheless facilitate a >5x amplification of a 74,000 bp region. In contrast to *pgi*, *tpiA* falls amidst several ribosomal RNA operons, and every single observed genome expansion was a duplication (2x read depth increase) resulting from homology-mediated amplification between different combinations of the rrn operons (Figure 5.5C). Taken together, the high frequency of *tpiA* amplifications in early midpoint strains and low frequency in endpoint strains implies that these amplifications have an ease-of-acquisition rather than fitness benefit, compared to other mechanisms of orthogene upregulation. This is supported by a 15% higher growth rate resulting from the *yiiQ* L179L SNP, as compared to a genome duplication.

Fortuitously, enough clones were isolated and sequenced to find 5 strains genetically identically except for extent of amplified region, allowing for phenotypic comparison (Figure 5.5D). Under glucose growth conditions, duplication of the middle (*tpiA*-containing), or middle + left regions leads to highest growth rates, while inclusion of the right region lowers the growth rate. Examining different growth environments, we

see drastic changes in relative fitness – on xylose growth all differences in strain fitness are essentially negated, while on glycerol growth the middle and middle + right duplications having significantly higher fitness than any duplication that includes the left region. Thus, specifics of local gene context can not only influence accessible adaptive mutations, but also facilitate (or hinder) expansion into entirely new niches.

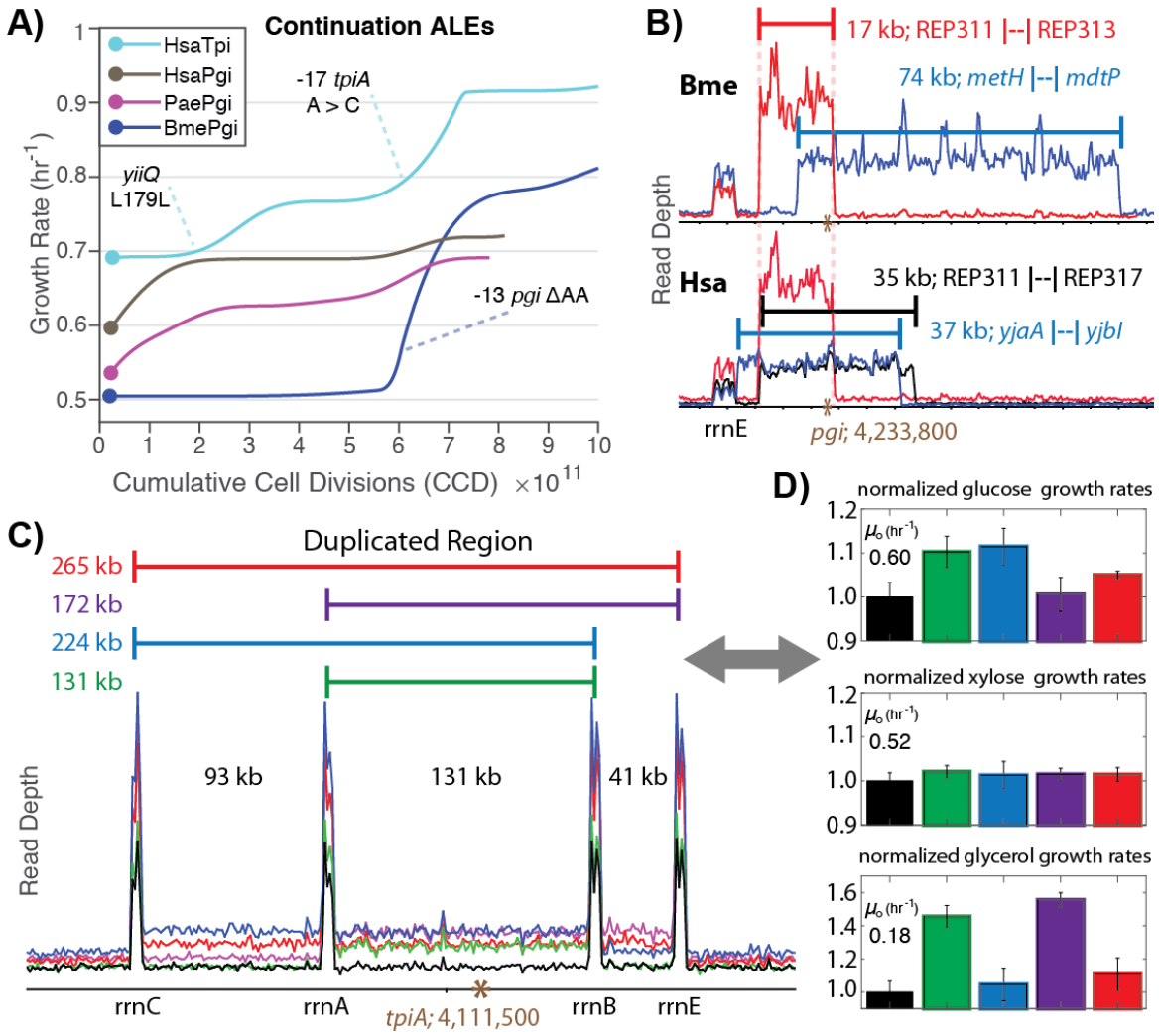



**Figure 5.5: Adaptive dynamics and genome rearrangements** (A) Continuation ALEs had the potential to enable escape from a 'failed' trajectory, even in the presence of multiple KO-characteristic mutations (B) All genome expansions found in *pgi* swap strains (C) All genome expansions found in *tpiA* swap strains (D) Relative growth rates on various carbon sources for HsaTpi strains genetically identical except for the size of *tpiA* duplication; coloration corresponds with (C)

### 5.2.6 Structural Elucidation of Mutation Hotspots

Across all strains, the genes with the highest number of independent mutational hits (*rpoB, rpoC, rpoA*) were components of the RNA polymerase complex, as is frequently the case in ALE experiments [36]. These mutations effect global transcriptional shifts that can significantly alter cell physiology and proteome composition, but detailed interpretation is notoriously difficult, often requiring whole studies to investigate single mutations [110]. In this study, our large sample size resulted in a plethora of observed mutations which, when taken together, point to several structural regions as key accumulators of adaptive mutations. Most notably, C-terminal *rpoA* mutations were a rampant mechanism of adaptation for both Δ*pgi* and failure *pgi* swap lineages – the *rpoA* G315V mutation was observed 5 times independently across various strains, along with several frameshifting mutations falling within the same 5 bp window. Contrastingly, *tpiA* swaps never acquired these C-terminal *rpoA* mutations, but instead targeted glycine residue 36 for alteration, with *rpoA* G36D, G36V, and G36C SNPs observed across one or more *tpiA* swap strains.

Although mutation hotspot regions can be deduced from mutational recurrence in protein primary structure, this overlooks potential hotspots involving SNPs distant in primary structure but localized in the same region of the full quaternary protein structure. Recent cryoEM techniques are yielding an unprecedented view into the *in vivo* dynamics of protein complexes, and mapping our 75+ unique observed *rpo* SNPs to the *E. coli* transcriptional activation complex [155] revealed distinct spatial clusters (Figure 5.6). This clustering highlighted mutational associations that would not have been drawn without such structural interpretation: the frequently observed C-terminal *rpoA* mutations

were found to fall in close proximity to the various *crp* mutations. Given that both mutation types are characteristic of Δ*pgi*, and fall directly in a protein-DNA interaction region, this implies a shared mechanism of phenotypic influence via altered transcription initiation dynamics. This is in contrast to the *rpoA* G36 mutations, which have close proximity to important *rpoC* SNPs, but no association with *crp* (Figure 5.6C)

A recent study by Wytock *et al.* examined adaptation following a range of metabolic gene KOs, which despite significant genetic and growth rate differences still converged on several instances of identical RNA polymerase mutations [156]. Here we see striking agreement between the Wytock mutations and many found in this study – *rpoB* residues D516, P520, D654, and *rpoC* residues R1075 and R978 all acquired multiple independent mutations across gene-swapped strains. Many of these most frequently targeted residues fell into distinct spatial clusters with one another, despite amino acid positions hundreds of residues distant in primary structure (Figure 5.6C). The spectrum and frequency of adaptive *rpo* mutations found with ALE essentially amounts to an empirical, rather than computational, Monte Carlo sampling of the 'solution space' for RNA polymerase alterations capable of enabling fast growth in spite of metabolic lesions. This sampling revealed the surprising ability for identical mutations to increase growth across a diverse range of metabolic perturbations, pointing to global principles underlying transcriptional dynamics and resultant cell physiology. Informed by ALE-derived sampling of structural solution spaces and increasingly available cryoEM structures, deciphering complex biochemical dynamics and their physiological influence may soon be feasible.

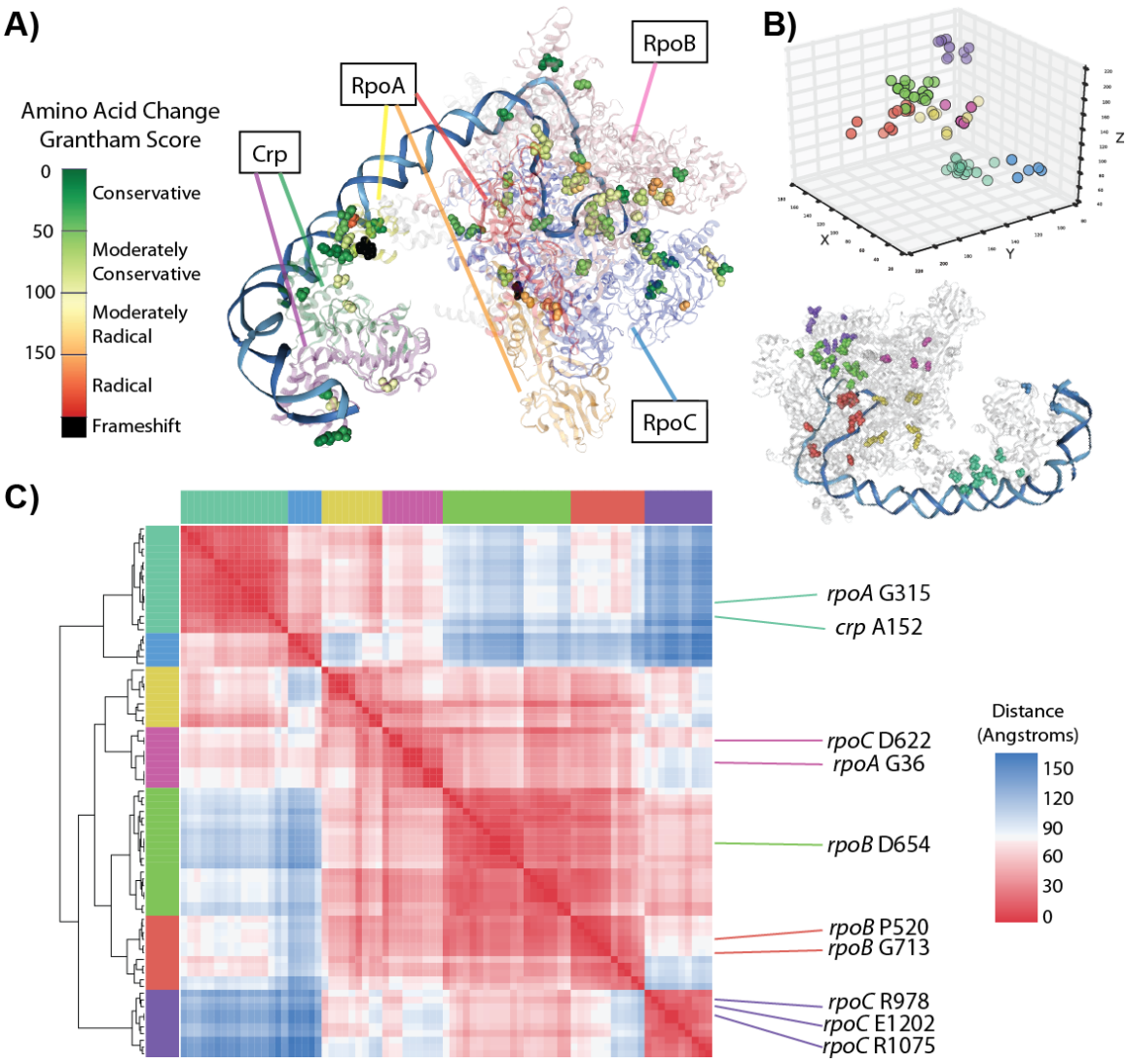



**Figure 5.6: Structural clustering of RNApol complex SNPs** (A) RNA Polymerase transcription activation complex cryoEM structure; observed mutations across all gene-swapped and KO strains colored based on amino acid replacement distance (B) Clustering of mutations identified hotspot regions in 3D space (C) Various hotspot regions were characteristic of repeatedly observed SNPs

## 5.3 Discussion

In this study, we genetically engineered *E. coli* to replace native copies of *pgi* and *tpiA* with foreign orthologs from all domains of life, and evolved dozens of independent lineages for hundreds to thousands of generations. We found that, despite initial impaired

or negligible functionality, strains readily evolved to 'assimilate' the foreign gene, reaching growth rates just as high as found in evolved wild-type *E. coli*. We examined the mutational mechanisms by which orthogenes were either successfully assimilated or relegated to disuse, uncovering specific features of both the DNA and protein sequences that became targets for adaptive mutations in the new *E. coli* host. With hundreds of genetically distinct evolved clones, we identified multiple genes and even specific protein residues that were repeated mutational targets. While some genes were mutated uniquely under specific conditions, others provided fitness benefits across a range of strain types, and structural analysis of RNA polymerase mutations pointed to specific spatial regions underlying adaptive shifts in global expression state. We established the contingency of accessible adaptive mutations on a gene's chromosomal location, and demonstrated how the specifics of local gene context can facilitate expansion into new niches.

Increasing orthogene expression was the key determinate of a strain's 'success' or 'failure' to assimilate the foreign DNA, and a variety of mutational mechanisms were observed. Predominantly, expression-increasing mutations did not alter the foreign DNA at all, but instead targeted cis-regulatory regions such as ribosomal and promoter binding sites. The gene chosen for swapping led to significantly different mutational outcomes, with promoter/RBS SNPs dominant for *pgi,* but *tpiA* swaps characterized predominantly by a silent SNP in the upstream gene that creates a new promoter. Similarly, chromosomal location constrained both the accessibility of genome amplifications as a method of upregulation, and cross-environment fitness changes stemming from the specific gene content of the amplified region. In addition to the influence of local sequence context, specifics of the swapped gene also altered adaptive trajectories. The

archael PaeTpi swap induced poly-proline ribosomal stalling that was ameliorated by missense SNPs to either of the two prolines, while HsaTpi and BmePgi swaps had prohibitively strong mRNA secondary structures that were targeted for destabilization by stem-loop SNPs.

With strongly demonstrated causal influence from multiple different synonymous mutations, both promoter-creating and mRNA-targeting, a reevaluation of the common assumption that synonymous SNPs are phenotypically neutral is in order, and thus the applicability of Ka/Ks in evolutionary biology. Indeed, calculations of adaptive protein substitution rates are known to be sensitive to even weak selection for synonymous mutations [157]. We also provided evidence that missense SNPs can sometimes induce phenotypic changes by targeting mRNA stem-loops for destabilization, rather than due to any alterations in enzyme activity caused by the new amino acid. In light of these findings, a reexamination of mutations observed in ALE experiments is warranted – adaptive mutations may have been overlooked if synonymous, or misinterpreted if residue-altering. Similarly, databases of genome sequences are ripe for computational analysis to identify the extent to which stem-loop SNPs play a role in adaptation across different gene types and organisms.

Our findings also have implications for the rational design of strains for heterologous production. Codon optimization is unnecessary to enable orthogene functionality, as demonstrated herein, and if done naïvely may in fact induce mRNA stem-loops that countervail any expression increases from improved codon adaptation index. The strong influence of N-terminal mRNA structure on expression levels suggests new techniques for expression tuning – pairing a strong RBS with intentional stem-loops

of varying strength could achieve a range of expression levels. This variation would be due to changes in flux/mRNA, which we saw has a fitness advantage over less parsimonious methods of expression increase, such as promoter mutations or genome amplifications. Also relevant for genetic engineering, specifically evolutionary stability, we found that genome placement strongly influences the nature and likelihood of both adaptive point mutations and genome amplifications.

Overall, this study establishes the impressive adaptive flexibility of *E. coli* - only a few mutations are necessary to get a gene working in a new host, despite billions of years of evolutionary divergence from the replaced native copies. With the rising use of gene KO ALEs to investigate metabolic perturbations and thus understand cellular functions in greater detail [156, 158], the results of this study imply similar fruitful investigations enabled by gene-swaps. Depending on donor organism specifics, such swaps can introduce new metabolic capabilities, probe genetic interchangeability, or serve as gene knock-downs rather than knock-outs. With a carefully selected experimental design for appropriate selection pressure, ALE can be used both to facilitate traditional methods of heterologous strain construction and make biological discoveries of pertinent protein/DNA features necessary for robust *in vivo* functionality.

## 5.4 Materials and Methods

### 5.4.1 Strain Design and Engineering

DNA sequences for gene replacement were ordered from Gene Universal Inc. For the human genes, the coding sequence (introns removed) of the annotated main isoform was used. Strains were constructed in two different ways - *pgi* swaps with a modified gene gorging protocol [159] as depicted in Figure 5.7, and *tpiA* swaps with a similar method but using CRISPR-induced double-stranded DNA breaks on the native *E. coli* sequence as the method of counter-selection, thus not requiring an antibiotic cassette. Strain construction was checked for compositional and locational accuracy with both whole-genome and Sanger sequencing.

Strain phylogeny (Figure 5.1B) was compared with the TimeTree tool [160]. Protein similarity scores (Table 1) were obtained using EMBOSS Needle pairwise sequence alignment [161]. Codon adaptation index (Table 1) was calculated with the tool from Biologics International Corp (https://www.biologicscorp.com/tools/CAICalculator).

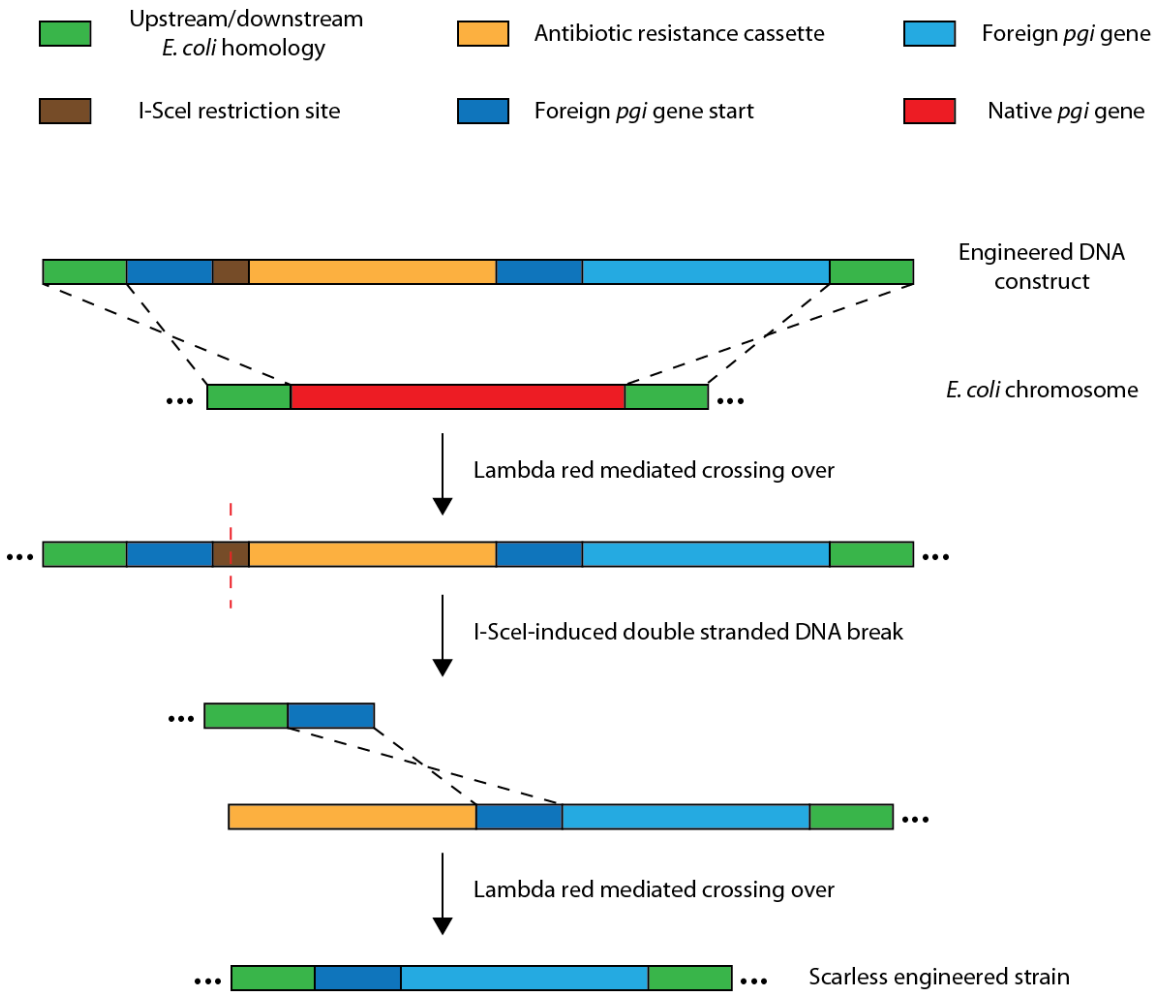



**Figure 5.7: Method of scarless strain construction.** The *pgi* swap strains were constructed as shown. The *tpiA* swap strains used a construct lacking the foreign gene start homology, I-SceI site, and antibiotic cassette; double stranded DNA breaks were induced by CRISPR-targeting to native gene sequence, obviating the need for antibiotics.

## 5.4.2 Adaptive Laboratory Evolution

Strains were evolved via batch culture serial propagation of 100 ul volumes into 15 mL (working volume) tubes of 4 g/L glucose M9 minimal media kept at 37 C and aerated via magnetic stirring, exactly as described previously [162]. Cultures were propagated for 30 days, or until measured population growth rate reached within 10% of the maximum known to occur for the growth conditions. Each independent ALE replicate

was started from a unique pre-culture inoculated with a clone isolated from an LB plate, so as to prevent standing variation from influencing mutational independence across replicates.

### 5.4.3 DNA Sequencing and Analysis

Genomic DNA was isolated using bead agitation in 96-well plates as outlined in [163]. Paired-end whole genome DNA sequencing libraries were generated with a Kapa HyperPlus library prep kit (Kapa Biosystems) and run on an Illumina HiSeq 4000 platform with a HiSeq SBS kit, 150 bp reads. The generated DNA sequencing fastq files were processed with the breseq computational pipeline [82] and aligned to the *E. coli* genome (NCBI accession NC_000913.3) to identify mutations. Genome amplifications were identified with a custom script, and all read depth coverage plots and marginal mutation calls were also manually inspected.

### 5.4.4 Strain Characterizations

Strains were assayed for growth rate via serial propagation and OD sampling conditions identical to the ALE experiments, with given values an average over at minimum 5 independent growth tubes. Strains were assayed for *pgi* expression level via RT-qPCR as follows: total RNA was purified with the Qiagen mRNeasy kit, assessed for quality on an Agilent Bioanalyzer, and quantified with a Nanodrop. RNA was converted to cDNA with the NEB LunaScript RT Kit, then quantified with the NEB Luna Universal qPCR kit. A panel of 5 housekeeping genes were assayed along with *pgi* to allow for normalization. Strains were assayed for enzymatic flux with a PGI activity colorimetric

assay kit from BioVision (#K775) as per manufacturer protocols. For both expression level and flux measurements, cultures were first flash frozen (both biological and technical duplicates) in liquid nitrogen at the same OD in mid-exponential growth phase.

### 5.4.5 Mutation Structural Analysis

Structures for mRNA transcripts were evaluated using the NUPACK computational tool [151]. Annotated transcription start sites served as the 5' end for evaluated structures, and a range of transcript lengths were analyzed to ensure robustness of results to region chosen. RNA Polymerase mutations were analyzed with a hierarchical clustering algorithm to identify the occurrence of non-random mutational "hotspots" on the transcriptional activation complex (PDB ID: 6b6h) from the 3D coordinates of the mutated amino acid residues. Mutations were clustered into n-groups, where $n \in [1, 18]$. For each number of clusters (n) selected, the mean-squared error (MSE) was identified using the formula below:

$$\sum_{j=1}^{n} \sum_{i,j} \left( D\left(C_j , P_{i,j}\right) \right)^2$$

Where the distance (D) between a cluster centroid ($C_j$) and each of the mutations identified to belong to that cluster ($P_{i,j}$) is squared and summed across the n-clusters. To quantify the "randomness" of the 3-dimensional localization of mutations, the MSE of n-clusters was compared to the MSE of hierarchical clustering in a random sample. Here, a random sample of amino acid positions was selected so that (1) the number of amino acids reflects the number of mutations in the data (Random or Un-Weighted Sampling); (2) the redundancy—number of repeated positions—of the mutation data is enforced on

the random selection (Enzyme-Weighted Sampling); and (3) the redundancy of the mutation data is enforced and restricted to each protein subunit (Chain-Weighted Sampling). The third sampling technique ensures that the distribution across protein subunits of amino acids selected matches that of the mutation data, and removes any bias of mutation occurrence across different protein subunits. At each number of clusters, the MSE of 30,000 random samples in each sampling technique was calculated. The mean and standard deviation of the distribution was used to derive p-values for the likelihood of the mutation clustering occurring by random chance, establishing statistical significance for the 7-cluster classification being both optimal and reflective of non-randomness.

Chapter 5 is a version of a manuscript in preparation for submission: Sandberg TE, Catoiu E, Szubin R, Phaneuf PV, Feist AM, Palsson BO. "Synthetic Cross-Domain Gene Replacement and Evolutionary Assimilation of Major Glycolytic Enzymes into *E. coli*." The dissertation author was the primary author of the paper and was responsible for the research.

# Chapter 6

# Conclusions and Outlook

In this dissertation, we performed in-depth analysis of the mechanisms by which *Escherichia coli* evolved to tolerate a diverse range of conditions. By studying perturbations of fundamentally different natures, we gained insight into both general and condition-specific adaptive strategies. Informed by a wide array of data types and physiological assays, we examined evolved strain phenotypes and linked their properties to the causal influence of various mutations. The lessons learned from such analyses, and implied future directions of research investigation, are discussed below.

## 6.1 Lessons Learned

In our investigation of adaptation to elevated temperature, we demonstrated the incredible sensitivity of mutational outcomes to both environmental particulars and starting genotype, established Multiplex Automated Genome Engineering as a method to prove mutational causality, and studied how global expression patterns shift over the course of adaptation. This highlighted the importance of appropriate control lineages for unbiased mutational comparison, and of independent ALE replicates for easy causal mutation identification without the need for laborious genetic engineering. We saw that a strongly driving force for the adaptive process is a restoration of cellular physiology to

pre-perturbed levels. This preferred homeostatic state could be reacquired with a minimal number of mutation steps due to the large-scale expression changes wrought by regulatory or RNA polymerase mutations. Though increasing fitness overall, we found that such mutations can also carry along actively detrimental expression changes, which subsequently become targets for additional ameliorating mutations.

In our investigation of adaptation to alternate growth substrates, we established the impressive ability of *E. coli* to exhibit completely normal physiology even when every carbon atom composing the cell is Carbon-13 rather than Carbon-12. Given the importance of $^{13}$C-labelling in metabolic flux analysis experiments, and previous questions on the validity of neglecting any potential kinetic isotope effect, this result is a sigh of relief to metabolic engineers. Examining a more complex carbon growth substrate perturbation, we added temporal variability to the typically-constant ALE conditions, investigating how *E. coli* adapted to rapidly alternating growth nutrients. We showed that the particulars of the alternating environment led to either co-existing specialist subpopulations or the predominance of a single generalist strain. Informed by these results, we then established the utility of computational techniques, specifically genome-scale metabolic modeling, to interpret or even predict such condition-specific evolutionary outcomes.

In our investigation of adaptation to genetic perturbations by foreign DNA, we demonstrated the remarkable ability of *E. coli* to tolerate replacement of important metabolic genes with orthologous copies that have been evolutionarily diverging for billions of years. We showed how, in addition to cis-regulatory elements, specific DNA or protein features were targeted for mutations that enabled increased orthogene

expression, frequently resulting in evolved growth rates of the same level achievable by the wild-type. We established genes and even structural regions that were repeated mutational targets either across conditions or specific to particular ones, and revealed the contingency of evolutionary outcomes on the genomic location of mutational hotspots. We also provided both observational and empirical evidence for the adaptive importance of mRNA stem-loop SNPs, thus partially resolving questions on the possible mechanism of phenotypic improvement for synonymous mutations.

## 6.2 The Future of ALE

As demonstrated herein, Adaptive Laboratory Evolution experiments can be a powerful method to probe living systems. Among the myriad variables and molecular interactions that come together to form a cohesive living cell, ALE is able to consistently and reproducibly point to the few genes or cellular processes most strongly under selection in a given environment. It thus serves as a tool for biological discovery (by elucidating adaptive mechanisms), strain engineering (by generating desired phenotypes), and hypothesis generation (by pointing to subtle effects that indicate potentially novel causal influences).

Key to the successful use of ALE throughout this dissertation were the automated 'ALE machine' platforms, enabling experimental control, throughput, and data quality at heretofore unachievable levels. The increased capabilities bestowed by automation are leading others to develop similar systems [164, 165], and ALE is thus poised to see even wider adoption. As the number of ALE studies grows their utility will only increase further – the preexisting datasets serve as essential comparison points, allowing

discernment between mutations specific to a perturbation of interest or merely some other factor. Databases of such ALE-identified mutations are ripe for computational probing using big data analytics techniques to discover evolutionary features hidden in the massive multidimensionality of genotype space [166].

Together with automation, advances in genetic engineering and synthetic biology open up exciting new possibilities for ALE. With synthetic gene circuits begging to see *in vivo* therapeutic use [167], ALE provides a way to both examine and engineer the evolutionary stability of such systems. Ambitious alterations such as genomic recoding leave microbes with severe growth defects, and ALE allows for optimization of these synthetic strains [168, 169]. New CRISPR-based tools for targeted mutagenesis [170] provide the ability to force previously stochastic adaptive walks down desired paths. Taken all together, the outlook for ALE appears bright – evolutionary engineering allows for optimization of dauntingly complex systems, and evolutionary investigations open new lines of inquiry into biological phenomena.

# References


1. Chen X, Zhou L, Tian K, Kumar A, Singh S, Prior BA, Wang Z. Metabolic engineering of Escherichia coli: a sustainable industrial platform for bio-based chemical production. Biotechnology advances. 2013;31(8):1200-23. doi: 10.1016/j.biotechadv.2013.02.009. PubMed PMID: 23473968.

2. Dragosits M, Mattanovich D. Adaptive laboratory evolution -- principles and applications for biotechnology. Microb Cell Fact. 2013;12:64. Epub 2013/07/03. doi: 10.1186/1475-2859-12-64. PubMed PMID: 23815749; PubMed Central PMCID: PMC3716822.

3. Gresham D, Dunham MJ. The enduring utility of continuous culturing in experimental evolution. Genomics. 2014;104(6 Pt A):399-405. doi: 10.1016/j.ygeno.2014.09.015. PubMed PMID: 25281774; PubMed Central PMCID: PMCPMC4411559.

4. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-Term Experimental Evolution in Escherichia-Coli .1. Adaptation and Divergence during 2,000 Generations. American Naturalist. 1991;138(6):1315-41. doi: Doi 10.1086/285289. PubMed PMID: ISI:A1991HB55000001.

5. 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. doi: 10.1128/AEM.02246-14. PubMed PMID: 25304508; PubMed Central PMCID: PMC4272732.

6. Carroll SM, Marx CJ. Evolution after introduction of a novel metabolic pathway consistently leads to restoration of wild-type physiology. PLoS Genet. 2013;9(4):e1003427. Epub 2013/04/18. doi: 10.1371/journal.pgen.1003427. PubMed PMID: 23593025; PubMed Central PMCID: PMC3616920.

7. Cheng KK, Lee BS, Masuda T, Ito T, Ikeda K, Hirayama A, Deng L, Dong J, Shimizu K, Soga T, Tomita M, Palsson BO, Robert M. Global metabolic network reorganization by adaptive mutations allows fast growth of Escherichia coli on glycerol. Nat Commun. 2014;5:3233. Epub 2014/02/01. doi: 10.1038/ncomms4233. PubMed PMID: 24481126.

8. Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual populations. Science. 2013;342(6164):1364-7. doi: 10.1126/science.1243357. PubMed PMID: 24231808.

9. Portnoy VA, Bezdan D, Zengler K. Adaptive laboratory evolution--harnessing the power of biology for metabolic engineering. Curr Opin Biotechnol. 2011;22(4):590-4. Epub 2011/04/19. doi: 10.1016/j.copbio.2011.03.007. PubMed PMID: 21497080.

10. Ibarra RU, Edwards JS, Palsson BO. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002;420(6912):186-9. Epub 2002/11/15. doi: 10.1038/nature01149. PubMed PMID: 12432395.

11. Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BO. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet. 2010;6(11):e1001186. Epub 2010/11/17. doi: 10.1371/journal.pgen.1001186. PubMed PMID: 21079674; PubMed Central PMCID: PMC2973815.

12. Stoebel DM, Hokamp K, Last MS, Dorman CJ. Compensatory evolution of gene regulation in response to stress by Escherichia coli lacking RpoS. PLoS Genet. 2009;5(10):e1000671. Epub 2009/10/03. doi: 10.1371/journal.pgen.1000671. PubMed PMID: 19798444; PubMed Central PMCID: PMC2744996.

13. Goodarzi H, Bennett BD, Amini S, Reaves ML, Hottes AK, Rabinowitz JD, Tavazoie S. Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in E. coli. Mol Syst Biol. 2010;6:378. Epub 2010/06/10. doi: 10.1038/msb.2010.33. PubMed PMID: 20531407; PubMed Central PMCID: PMC2913397.

14. Rao VSH, Rao PRS. Global stability in chemostat models involving time delays and wall growth. Nonlinear Anal-Real. 2004;5(1):141-58. doi: Doi 10.1016/S1468-1218(03)00022-1. PubMed PMID: ISI:000185721300008.

15. Rozen DE, Philippe N, Arjan de Visser J, Lenski RE, Schneider D. Death and cannibalism in a seasonal environment facilitate bacterial coexistence. Ecol Lett. 2009;12(1):34-44. Epub 2008/11/21. doi: 10.1111/j.1461-0248.2008.01257.x. PubMed PMID: 19019196.

16. Wood TE, Burke JM, Rieseberg LH. Parallel genotypic adaptation: when evolution repeats itself. Genetica. 2005;123(1-2):157-70. Epub 2005/05/11. PubMed PMID: 15881688; PubMed Central PMCID: PMC2442917.

17. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. Programming cells by multiplex genome engineering and accelerated evolution. Nature. 2009;460(7257):894-8. Epub 2009/07/28. doi: 10.1038/nature08187. PubMed PMID: 19633652.

18. Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. The molecular diversity of adaptive convergence. Science. 2012;335(6067):457-61. Epub 2012/01/28. doi: 10.1126/science.1212986. PubMed PMID: 22282810.

19. Fong SS, Joyce AR, Palsson BO. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res. 2005;15(10):1365-72. Epub 2005/10/06. doi: 10.1101/gr.3832305. PubMed PMID: 16204189; PubMed Central PMCID: PMC1240078.

20. Camps M, Naukkarinen J, Johnson BP, Loeb LA. Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. Proc Natl Acad Sci U S A. 2003;100(17):9727-32. Epub 2003/08/12. doi: 10.1073/pnas.1333928100. PubMed PMID: 12909725; PubMed Central PMCID: PMC187833.

21. Lee DH, Feist AM, Barrett CL, Palsson BO. Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli. PLoS One. 2011;6(10):e26172. doi: 10.1371/journal.pone.0026172. PubMed PMID: 22028828; PubMed Central PMCID: PMC3196513.

22. Lee HH, Molla MN, Cantor CR, Collins JJ. Bacterial charity work leads to population-wide resistance. Nature. 2010;467(7311):82-5. doi: 10.1038/nature09354. PubMed PMID: 20811456; PubMed Central PMCID: PMC2936489.

23. Shachrai I, Zaslaver A, Alon U, Dekel E. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. Mol Cell. 2010;38(5):758-67. Epub 2010/05/04. doi: 10.1016/j.molcel.2010.04.015. PubMed PMID: 20434381.

24. Vemuri GN, Altman E, Sangurdekar DP, Khodursky AB, Eiteman MA. Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio. Appl Environ Microbiol. 2006;72(5):3653-61. Epub 2006/05/05. doi: 10.1128/AEM.72.5.3653-3661.2006. PubMed PMID: 16672514; PubMed Central PMCID: PMC1472329.

25. Sniegowski PD, Gerrish PJ, Lenski RE. Evolution of high mutation rates in experimental populations of E. coli. Nature. 1997;387(6634):703-5. Epub 1997/06/12. doi: 10.1038/42701. PubMed PMID: 9192894.

26. Ban C, Yang W. Crystal structure and ATPase activity of MutL: implications for DNA repair and mutagenesis. Cell. 1998;95(4):541-52. Epub 1998/11/25. PubMed PMID: 9827806.

27. Echols H, Lu C, Burgers PM. Mutator strains of Escherichia coli, mutD and dnaQ, with defective exonucleolytic editing by DNA polymerase III holoenzyme. Proc Natl Acad Sci U S A. 1983;80(8):2189-92. Epub 1983/04/01. PubMed PMID: 6340117; PubMed Central PMCID: PMC393783.

28. Conrad TM, Joyce AR, Applebee MK, Barrett CL, Xie B, Gao Y, Palsson BO. 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. 2009;10(10):R118. Epub 2009/10/24. doi: 10.1186/gb-2009-10-10-r118
gb-2009-10-10-r118 [pii]. PubMed PMID: 19849850; PubMed Central PMCID: PMC2784333.

29. Blank D, Wolf L, Ackermann M, Silander OK. The predictability of molecular evolution during functional innovation. Proc Natl Acad Sci U S A. 2014;111(8):3044-9. Epub 2014/02/12. doi: 10.1073/pnas.1318797111. PubMed PMID: 24516157.

30. Jensen KF. 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. 1993;175(11):3401-7. Epub 1993/06/01. PubMed PMID: 8501045; PubMed Central PMCID: PMC204738.

31. UniProt. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42(Database issue):D191-8. Epub 2013/11/21. doi: 10.1093/nar/gkt1140. PubMed PMID: 24253303.

32. Kishimoto T, Iijima L, Tatsumi M, Ono N, Oyake A, Hashimoto T, Matsuo M, Okubo M, Suzuki S, Mori K, Kashiwagi A, Furusawa C, Ying BW, Yomo T. Transition from positive to neutral in mutation fixation along with continuing rising fitness in thermal adaptive evolution. PLoS Genet. 2010;6(10):e1001164. Epub 2010/10/27. doi: 10.1371/journal.pgen.1001164. PubMed PMID: 20975944; PubMed Central PMCID: PMC2958811.

33. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc Natl Acad Sci U S A. 2006;103(24):9107-12. Epub 2006/06/06. doi: 10.1073/pnas.0602917103. PubMed PMID: 16751270; PubMed Central PMCID: PMC1482574.

34. Klein-Marcuschamer D, Santos CN, Yu H, Stephanopoulos G. Mutagenesis of the bacterial RNA polymerase alpha subunit for improvement of complex phenotypes. Appl Environ Microbiol. 2009;75(9):2705-11. Epub 2009/03/03. doi: 10.1128/AEM.01888-08. PubMed PMID: 19251886; PubMed Central PMCID: PMC2681691.

35. Murphy H, Cashel M. Isolation of RNA polymerase suppressors of a (p)ppGpp deficiency. Methods Enzymol. 2003;371:596-601. Epub 2004/01/10. doi: 10.1016/S0076-6879(03)71044-1. PubMed PMID: 14712731.

36. Conrad TM, Frazier M, Joyce AR, Cho BK, Knight EM, Lewis NE, Landick R, Palsson BO. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. Proc Natl Acad Sci U S A. 2010;107(47):20500-5. Epub 2010/11/09. doi: 10.1073/pnas.0911253107. PubMed PMID: 21057108; PubMed Central PMCID: PMC2996682.

37. Jain C, Deana A, Belasco JG. Consequences of RNase E scarcity in Escherichia coli. Mol Microbiol. 2002;43(4):1053-64. Epub 2002/04/04. PubMed PMID: 11929550.

38. Perwez T, Hami D, Maples VF, Min Z, Wang BC, Kushner SR. Intragenic suppressors of temperature-sensitive rne mutations lead to the dissociation of RNase E activity on mRNA and tRNA substrates in Escherichia coli. Nucleic Acids Res. 2008;36(16):5306-18. Epub 2008/08/12. doi: 10.1093/nar/gkn476. PubMed PMID: 18689439; PubMed Central PMCID: PMC2532720.

39. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol.

2010;6(2):e1000664. Epub 2010/02/09. doi: 10.1371/journal.pcbi.1000664. PubMed PMID: 20140241; PubMed Central PMCID: PMC2816680.

40. Folichon M, Arluison V, Pellegrini O, Huntzinger E, Regnier P, Hajnsdorf E. The poly(A) binding protein Hfq protects RNA from RNase E and exoribonucleolytic degradation. Nucleic Acids Res. 2003;31(24):7302-10. Epub 2003/12/05. PubMed PMID: 14654705; PubMed Central PMCID: PMC291859.

41. Ringquist S, Shinedling S, Barrick D, Green L, Binkley J, Stormo GD, Gold L. Translation initiation in Escherichia coli: sequences within the ribosome-binding site. Mol Microbiol. 1992;6(9):1219-29. Epub 1992/05/01. PubMed PMID: 1375310.

42. Rodionov DA, Gelfand MS, Mironov AA, Rakhmaninova AB. Comparative approach to analysis of regulation in complete genomes: multidrug resistance systems in gamma-proteobacteria. J Mol Microbiol Biotechnol. 2001;3(2):319-24. Epub 2001/04/26. PubMed PMID: 11321589.

43. Piddock LJ. Multidrug-resistance efflux pumps - not just for resistance. Nat Rev Microbiol. 2006;4(8):629-36. Epub 2006/07/18. doi: 10.1038/nrmicro1464. PubMed PMID: 16845433.

44. Plumbridge J. An alternative route for recycling of N-acetylglucosamine from peptidoglycan involves the N-acetylglucosamine phosphotransferase system in Escherichia coli. J Bacteriol. 2009;191(18):5641-7. Epub 2009/07/21. doi: 10.1128/JB.00448-09. PubMed PMID: 19617367; PubMed Central PMCID: PMC2737974.

45. Vijayendran C, Barsch A, Friehs K, Niehaus K, Becker A, Flaschel E. Perceiving molecular evolution processes in Escherichia coli by comprehensive metabolite and gene expression profiling. Genome Biol. 2008;9(4):R72. Epub 2008/04/12. doi: 10.1186/gb-2008-9-4-r72. PubMed PMID: 18402659; PubMed Central PMCID: PMC2643943.

46. Sinensky M. Homeoviscous adaptation--a homeostatic process that regulates the viscosity of membrane lipids in Escherichia coli. Proc Natl Acad Sci U S A. 1974;71(2):522-5. Epub 1974/02/01. PubMed PMID: 4360948; PubMed Central PMCID: PMC388039.

47. Blaby IK, Lyons BJ, Wroclawska-Hughes E, Phillips GC, Pyle TP, Chamberlin SG, Benner SA, Lyons TJ, Crecy-Lagard V, Crecy E. Experimental evolution of a facultative thermophile from a mesophilic ancestor. Appl Environ Microbiol. 2012;78(1):144-55. Epub 2011/10/25. doi: 10.1128/AEM.05773-11. PubMed PMID: 22020511; PubMed Central PMCID: PMC3255606.

48. Riehle MM, Bennett AF, Long AD. Genetic architecture of thermal adaptation in Escherichia coli. Proc Natl Acad Sci U S A. 2001;98(2):525-30. Epub 2001/01/10. doi: 10.1073/pnas.021448998. PubMed PMID: 11149947; PubMed Central PMCID: PMC14620.

49. Michel B. Replication fork arrest and DNA recombination. Trends Biochem Sci. 2000;25(4):173-8. Epub 2000/04/08. PubMed PMID: 10754549.

50. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440-5. Epub 2003/07/29. doi: 10.1073/pnas.1530509100. PubMed PMID: 12883005; PubMed Central PMCID: PMC170937.

51. Gunasekera TS, Csonka LN, Paliy O. Genome-wide transcriptional responses of Escherichia coli K-12 to continuous osmotic and heat stresses. J Bacteriol. 2008;190(10):3712-20. Epub 2008/03/25. doi: 10.1128/JB.01990-07. PubMed PMID: 18359805; PubMed Central PMCID: PMC2395010.

52. Gyaneshwar P, Paliy O, McAuliffe J, Jones A, Jordan MI, Kustu S. Lessons from Escherichia coli genes similarly regulated in response to nitrogen and sulfur limitation. Proc Natl Acad Sci U S A. 2005;102(9):3453-8. doi: 10.1073/pnas.0500141102. PubMed PMID: 15716358; PubMed Central PMCID: PMC552917.

53. Chaulk SG, Smith Frieday MN, Arthur DC, Culham DE, Edwards RA, Soo P, Frost LS, Keates RA, Glover JN, Wood JM. ProQ is an RNA chaperone that controls ProP levels in Escherichia coli. Biochemistry. 2011;50(15):3095-106. Epub 2011/03/09. doi: 10.1021/bi101683a. PubMed PMID: 21381725.

54. Soutourina OA, Bertin PN. Regulation cascade of flagellar expression in Gram-negative bacteria. FEMS Microbiol Rev. 2003;27(4):505-23. Epub 2003/10/11. PubMed PMID: 14550943.

55. Cooper TF, Rozen DE, Lenski RE. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A. 2003;100(3):1072-7. Epub 2003/01/23. doi: 10.1073/pnas.0334340100. PubMed PMID: 12538876; PubMed Central PMCID: PMC298728.

56. Landini P, Zehnder AJ. The global regulatory hns gene negatively affects adhesion to solid surfaces by anaerobically grown Escherichia coli by modulating expression of flagellar genes and lipopolysaccharide production. J Bacteriol. 2002;184(6):1522-9. Epub 2002/03/02. PubMed PMID: 11872702; PubMed Central PMCID: PMC134881.

57. Wang L, Wang FF, Qian W. Evolutionary rewiring and reprogramming of bacterial transcription regulation. Journal of genetics and genomics. 2011;38(7):279-88. Epub 2011/07/23. doi: 10.1016/j.jgg.2011.06.001. PubMed PMID: 21777852.

58. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009;461(7268):1243-7. Epub 2009/10/20. doi: 10.1038/nature08480. PubMed PMID: 19838166.

59. Duran-Figueroa NV, Pina-Escobedo A, Schroeder I, Simons RW, Garcia-Mena J. Polynucleotide phosphorylase interacts with ribonuclease E through a betabetaalphabetabetaalpha domain. Biochimie. 2006;88(6):725-35. Epub 2006/02/18. doi: 10.1016/j.biochi.2006.01.002. PubMed PMID: 16483707.

60. Rodriguez-Verdugo A, Gaut BS, Tenaillon O. Evolution of Escherichia coli rifampicin resistance in an antibiotic-free environment during thermal stress. BMC Evol Biol. 2013;13:50. doi: 10.1186/1471-2148-13-50. PubMed PMID: 23433244; PubMed Central PMCID: PMC3598500.

61. Navarro Llorens JM, Tormo A, Martinez-Garcia E. Stationary phase in gram-negative bacteria. FEMS Microbiol Rev. 2010;34(4):476-95. Epub 2010/03/20. doi: 10.1111/j.1574-6976.2010.00213.x. PubMed PMID: 20236330.

62. Hindre T, Knibbe C, Beslon G, Schneider D. New insights into bacterial adaptation through in vivo and in silico experimental evolution. Nat Rev Microbiol. 2012;10(5):352-65. doi: 10.1038/nrmicro2750. PubMed PMID: 22450379.

63. Harano Y, Kinoshita M. Translational-entropy gain of solvent upon protein folding. Biophys J. 2005;89(4):2701-10. Epub 2005/08/02. doi: 10.1529/biophysj.104.057604. PubMed PMID: 16055541; PubMed Central PMCID: PMC1366771.

64. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet. 2014;15(2):107-20. Epub 2014/01/17. doi: 10.1038/nrg3643. PubMed PMID: 24430943.

65. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMC3322381.

66. Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee DH, Qiu Y, Zengler K. The genome organization of Thermotoga maritima reflects its lifestyle. PLoS Genet. 2013;9(4):e1003485. Epub 2013/05/03. doi: 10.1371/journal.pgen.1003485 PGENETICS-D-12-01486 [pii]. PubMed PMID: 23637642; PubMed Central PMCID: PMC3636130.

67. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010;28(5):511-5. doi: 10.1038/nbt.1621. PubMed PMID: 20436464; PubMed Central PMCID: PMC3146043.

68. Antoniewicz MR. Methods and advances in metabolic flux analysis: a mini-review. J Ind Microbiol Biotechnol. 2015;42(3):317-25. doi: 10.1007/s10295-015-1585-x. PubMed PMID: 25613286.

69. Antoniewicz MR. Dynamic metabolic flux analysis--tools for probing transient states of metabolic networks. Curr Opin Biotechnol. 2013;24(6):973-8. doi: 10.1016/j.copbio.2013.03.018. PubMed PMID: 23611566.

70. Boghigian BA, Seth G, Kiss R, Pfeifer BA. Metabolic flux analysis and pharmaceutical production. Metabolic engineering. 2010;12(2):81-95. doi: 10.1016/j.ymben.2009.10.004. PubMed PMID: 19861167.

71. Klein S, Heinzle E. Isotope labeling experiments in metabolomics and fluxomics. Wiley interdisciplinary reviews Systems biology and medicine. 2012;4(3):261-72. doi: 10.1002/wsbm.1167. PubMed PMID: 22447740.

72. Farquhar GD, Ehleringer JR, Hubick KT. Carbon Isotope Discrimination and Photosynthesis. Annu Rev Plant Phys. 1989;40:503-37. doi: Doi 10.1146/Annurev.Arplant.40.1.503. PubMed PMID: WOS:A1989U857900019.

73. Wasylenko TM, Stephanopoulos G. Kinetic isotope effects significantly influence intracellular metabolite (13) C labeling patterns and flux determination. Biotechnology journal. 2013;8(9):1080-9. doi: 10.1002/biot.201200276. PubMed PMID: 23828762; PubMed Central PMCID: PMC4045492.

74. 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. Molecular biology and evolution. 2014;31(10):2647-62. doi: 10.1093/molbev/msu209. PubMed PMID: 25015645; PubMed Central PMCID: PMC4166923.

75. Zhou J, Rudd KE. EcoGene 3.0. Nucleic Acids Res. 2013;41(Database issue):D613-24. doi: 10.1093/nar/gks1235. PubMed PMID: 23197660; PubMed Central PMCID: PMC3531124.

76. Dragosits M, Mozhayskiy V, Quinones-Soto S, Park J, Tagkopoulos I. Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in Escherichia coli. Mol Syst Biol. 2013;9:643. doi: 10.1038/msb.2012.76. PubMed PMID: 23385483; PubMed Central PMCID: PMC3588905.

77. Torres-Cruz J, van der Woude MW. Slipped-strand mispairing can function as a phase variation mechanism in Escherichia coli. J Bacteriol. 2003;185(23):6990-4. PubMed PMID: 14617664; PubMed Central PMCID: PMC262711.

78. Moxon R, Bayliss C, Hood D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annual review of genetics. 2006;40:307-33. doi: 10.1146/annurev.genet.40.110405.090442. PubMed PMID: 17094739.

79. Hua Q, Yang C, Oshima T, Mori H, Shimizu K. Analysis of gene expression in Escherichia coli in response to changes of growth-limiting nutrient in chemostat cultures. Appl Environ Microbiol. 2004;70(4):2354-66. PubMed PMID: 15066832; PubMed Central PMCID: PMCPMC383082.

80. Crown SB, Long CP, Antoniewicz MR. Integrated 13C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. Metab Eng. 2015;28:151-8. doi: 10.1016/j.ymben.2015.01.001. PubMed PMID: 25596508.

81. Millard P, Portais JC, Mendes P. Impact of kinetic isotope effects in isotopic studies of metabolic systems. BMC Syst Biol. 2015;9(1):64. doi: 10.1186/s12918-015-0213-8. PubMed PMID: 26410690; PubMed Central PMCID: PMCPMC4583766.

82. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods in molecular biology. 2014;1151:165-88. doi: 10.1007/978-1-4939-0554-6_12. PubMed PMID: 24838886; PubMed Central PMCID: PMC4239701.

83. Au J, Choi J, Jones SW, Venkataramanan KP, Antoniewicz MR. Parallel labeling experiments validate Clostridium acetobutylicum metabolic network model for C metabolic flux analysis. Metab Eng. 2014;26:23-33. PubMed PMID: 25183671.

84. He L, Xiao Y, Gebreselassie N, Zhang F, Antoniewicz MR, Tang YJ, Peng L. Central metabolic responses to the overproduction of fatty acids in Escherichia coli based on 13C-metabolic flux analysis. Biotechnol Bioeng. 2014;111(3):575-85. doi: 10.1002/bit.25124. PubMed PMID: 24122357.

85. Antoniewicz MR, Kraynie DF, Laffend LA, Gonzalez-Lergier J, Kelleher JK, Stephanopoulos G. Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol. Metab Eng. 2007;9(3):277-92. doi: 10.1016/j.ymben.2007.01.003. PubMed PMID: 17400499; PubMed Central PMCID: PMC2048574.

86. Long CP, Antoniewicz MR. Quantifying biomass composition by gas chromatography/mass spectrometry. Anal Chem. 2014;86(19):9423-7. doi: 10.1021/ac502734e. PubMed PMID: 25208224.

87. Antoniewicz MR, Kelleher JK, Stephanopoulos G. Measuring deuterium enrichment of glucose hydrogen atoms by gas chromatography/mass spectrometry. Anal Chem. 2011;83(8):3211-6. doi: 10.1021/ac200012p. PubMed PMID: 21413777; PubMed Central PMCID: PMC3562696.

88. Fernandez CA, Des Rosiers C, Previs SF, David F, Brunengraber H. Correction of 13C mass isotopomer distributions for natural stable isotope abundance. J Mass Spectrom. 1996;31(3):255-62. Epub 1996/03/01. doi: 10.1002/(SICI)1096-9888(199603)31:3<255::AID-JMS290>3.0.CO;2-3 [pii] 10.1002/(SICI)1096-9888(199603)31:3<255::AID-JMS290>3.0.CO;2-3. PubMed PMID: 8799277.

89. Noor E, Eden E, Milo R, Alon U. Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell. 2010;39(5):809-20. doi: 10.1016/j.molcel.2010.08.031. PubMed PMID: 20832731.

90. Gorke B, Stulke J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nat Rev Microbiol. 2008;6(8):613-24. doi: 10.1038/nrmicro1932. PubMed PMID: 18628769.

91. Hempfling WP, Mainzer SE. Effects of varying the carbon source limiting growth on yield and maintenance characteristics of Escherichia coli in continuous culture. J Bacteriol. 1975;123(3):1076-87. PubMed PMID: 169226; PubMed Central PMCID: PMCPMC235832.

92. Savageau MA. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of Escherichia coli. Genetics. 1998;149(4):1677-91. PubMed PMID: 9691028; PubMed Central PMCID: PMCPMC1460280.

93. Zaldivar J, Nielsen J, Olsson L. Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. Applied microbiology and biotechnology. 2001;56(1-2):17-34. PubMed PMID: 11499926.

94. Siegal ML. Shifting sugars and shifting paradigms. PLoS Biol. 2015;13(2):e1002068. doi: 10.1371/journal.pbio.1002068. PubMed PMID: 25688600; PubMed Central PMCID: PMCPMC4331491.

95. Solopova A, van Gestel J, Weissing FJ, Bachmann H, Teusink B, Kok J, Kuipers OP. Bet-hedging during bacterial diauxic shift. Proc Natl Acad Sci U S A. 2014;111(20):7427-32. doi: 10.1073/pnas.1320063111. PubMed PMID: 24799698; PubMed Central PMCID: PMCPMC4034238.

96. New AM, Cerulus B, Govers SK, Perez-Samper G, Zhu B, Boogmans S, Xavier JB, Verstrepen KJ. Different levels of catabolite repression optimize growth in stable and variable environments. PLoS Biol. 2014;12(1):e1001764. doi: 10.1371/journal.pbio.1001764. PubMed PMID: 24453942; PubMed Central PMCID: PMCPMC3891604.

97. Bennett AF, Lenski RE, Mittler JE. Evolutionary Adaptation to Temperature .1. Fitness Responses of Escherichia-Coli to Changes in Its Thermal Environment. Evolution. 1992;46(1):16-30. doi: Doi 10.2307/2409801. PubMed PMID: WOS:A1992HD99300002.

98. Ketola T, Mikonranta L, Zhang J, Saarinen K, Ormala AM, Friman VP, Mappes J, Laakso J. Fluctuating temperature leads to evolution of thermal generalism and preadaptation to novel environments. Evolution. 2013;67(10):2936-44. doi: 10.1111/evo.12148. PubMed PMID: 24094344.

99. Hughes BS, Cullum AJ, Bennett AF. An experimental evolutionary study on adaptation to temporally fluctuating pH in Escherichia coli. Physiol Biochem Zool. 2007;80(4):406-21. doi: 10.1086/518353. PubMed PMID: 17508336.

100. Alcantara-Diaz D, Brena-Valle M, Serment-Guerrero J. Divergent adaptation of Escherichia coli to cyclic ultraviolet light exposures. Mutagenesis. 2004;19(5):349-54. doi: 10.1093/mutage/geh039. PubMed PMID: 15388806.

101. Karve SM, Daniel S, Chavhan YD, Anand A, Kharola SS, Dey S. Escherichia coli populations in unpredictably fluctuating environments evolve to face novel stresses through enhanced efflux activity. J Evol Biol. 2015;28(5):1131-43. doi: 10.1111/jeb.12640. PubMed PMID: 25865653.

102. Kassen R. The experimental evolution of specialists, generalists, and the maintenance of diversity. J Evolution Biol. 2002;15(2):173-90. doi: DOI 10.1046/j.1420-9101.2002.00377.x. PubMed PMID: WOS:000174709000001.

103. Jasmin JN, Kassen R. Evolution of a single niche specialist in variable environments. Proc Biol Sci. 2007;274(1626):2761-7. doi: 10.1098/rspb.2007.0936. PubMed PMID: 17725975; PubMed Central PMCID: PMCPMC2279222.

104. Saxer G, Doebeli M, Travisano M. The repeatability of adaptive radiation during long-term experimental evolution of Escherichia coli in a multiple nutrient environment. PLoS One. 2010;5(12):e14184. doi: 10.1371/journal.pone.0014184. PubMed PMID: 21152028; PubMed Central PMCID: PMCPMC2996281.

105. Friesen ML, Saxer G, Travisano M, Doebeli M. Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in Escherichia coli. Evolution. 2004;58(2):245-60. doi: DOI 10.1111/j.0014-3820.2004.tb01642.x. PubMed PMID: WOS:000220194200004.

106. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, Dahan O, Pilpel Y. Adaptive prediction of environmental changes by microorganisms. Nature. 2009;460(7252):220-4. doi: 10.1038/nature08112. PubMed PMID: 19536156.

107. Cooper TF, Lenski RE. Experimental evolution with E. coli in diverse resource environments. I. Fluctuating environments promote divergence of replicate populations. BMC Evol Biol. 2010;10:11. doi: 10.1186/1471-2148-10-11. PubMed PMID: 20070898; PubMed Central PMCID: PMCPMC2827396.

108. Quan S, Ray JC, Kwota Z, Duong T, Balazsi G, Cooper TF, Monds RD. Adaptive evolution of the lactose utilization network in experimentally evolved populations of Escherichia coli. PLoS Genet. 2012;8(1):e1002444. doi: 10.1371/journal.pgen.1002444. PubMed PMID: 22253602; PubMed Central PMCID: PMCPMC3257284.

109. Kram KE, Geiger C, Ismail WM, Lee H, Tang H, Foster PL, Finkel SE. Adaptation of Escherichia coli to Long-Term Serial Passage in Complex Medium: Evidence of Parallel Evolution. mSystems. 2017;2(2). doi: 10.1128/mSystems.00192-16. PubMed PMID: 28289732; PubMed Central PMCID: PMCPMC5340864.

110. Utrilla J, O'Brien EJ, Chen K, McCloskey D, Cheung J, Wang H, Armenta-Medina D, Feist AM, Palsson BO. Global Rebalancing of Cellular Resources by

Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. Cell Syst. 2016;2(4):260-71. doi: 10.1016/j.cels.2016.04.003. PubMed PMID: 27135538; PubMed Central PMCID: PMCPMC4853925.

111. Shenhar Y, Biran D, Ron EZ. Resistance to environmental stress requires the RNA chaperones CspC and CspE. Environ Microbiol Rep. 2012;4(5):532-9. doi: 10.1111/j.1758-2229.2012.00358.x. PubMed PMID: 23760898.

112. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 2015;528(7580):99-104. doi: 10.1038/nature15765. PubMed PMID: 26632588; PubMed Central PMCID: PMCPMC4843128.

113. Enjalbert B, Cocaign-Bousquet M, Portais JC, Letisse F. Acetate Exposure Determines the Diauxic Behavior of Escherichia coli during the Glucose-Acetate Transition. J Bacteriol. 2015;197(19):3173-81. doi: 10.1128/JB.00128-15. PubMed PMID: 26216845; PubMed Central PMCID: PMCPMC4560281.

114. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. Science. 2005;307(5717):1962-5. doi: 10.1126/science.1106914. PubMed PMID: 15790856.

115. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009;7(2):129-43. doi: 10.1038/nrmicro1949. PubMed PMID: 19116616; PubMed Central PMCID: PMCPMC3119670.

116. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nature biotechnology. 2010;28(3):245-8. doi: 10.1038/nbt.1614. PubMed PMID: 20212490; PubMed Central PMCID: PMCPMC3108565.

117. Schellenberger J, Palsson BO. Use of randomized sampling for analysis of metabolic networks. J Biol Chem. 2009;284(9):5457-61. doi: 10.1074/jbc.R800048200. PubMed PMID: 18940807.

118. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Zengler K, Palsson BO. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun. 2012;3. doi: ARTN 929
10.1038/ncomms1928. PubMed PMID: WOS:000306995000001.

119. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BO. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013;9:693. doi: 10.1038/msb.2013.52. PubMed PMID: 24084808; PubMed Central PMCID: PMCPMC3817402.

120. Slininger PJ, Thompson SR, Weber S, Liu ZL, Moon J. Repression of xylose-specific enzymes by ethanol in Scheffersomyces (Pichia) stipitis and utility of repitching

xylose-grown populations to eliminate diauxic lag. Biotechnol Bioeng. 2011;108(8):1801-15. doi: 10.1002/bit.23119. PubMed PMID: 21370229.

121. Mundhada H, Seoane JM, Schneider K, Koza A, Christensen HB, Klein T, Phaneuf PV, Herrgard M, Feist AM, Nielsen AT. Increased production of L-serine in Escherichia coli through Adaptive Laboratory Evolution. Metabolic engineering. 2017;39:141-50. doi: 10.1016/j.ymben.2016.11.008. PubMed PMID: 27908688.

122. Xia T, Eiteman MA, Altman E. Simultaneous utilization of glucose, xylose and arabinose in the presence of acetate by a consortium of Escherichia coli strains. Microb Cell Fact. 2012;11:77. doi: 10.1186/1475-2859-11-77. PubMed PMID: 22691294; PubMed Central PMCID: PMCPMC3514249.

123. Erbilgin O, Bowen BP, Kosina SM, Jenkins S, Lau RK, Northen TR. Dynamic substrate preferences predict metabolic properties of a simple microbial consortium. BMC Bioinformatics. 2017;18(1):57. doi: 10.1186/s12859-017-1478-2. PubMed PMID: 28114881; PubMed Central PMCID: PMCPMC5259839.

124. Roemhild R, Barbosa C, Beardmore RE, Jansen G, Schulenburg H. Temporal variation in antibiotic environments slows down resistance evolution in pathogenic Pseudomonas aeruginosa. Evol Appl. 2015;8(10):945-55. doi: 10.1111/eva.12330. PubMed PMID: 26640520; PubMed Central PMCID: PMCPMC4662347.

125. Murima P, McKinney JD, Pethe K. Targeting bacterial central metabolism for drug development. Chem Biol. 2014;21(11):1423-32. doi: 10.1016/j.chembiol.2014.08.020. PubMed PMID: 25442374.

126. Sandberg TE, Long CP, Gonzalez JE, Feist AM, Antoniewicz MR, Palsson BO. Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology. PLoS One. 2016;11(3):e0151130. doi: 10.1371/journal.pone.0151130. PubMed PMID: 26964043; PubMed Central PMCID: PMCPMC4786092.

127. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology. 2013;31(1):46-53. doi: 10.1038/nbt.2450. PubMed PMID: 23222703; PubMed Central PMCID: PMCPMC3869392.

128. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. Mol Syst Biol. 2011;7:535. doi: 10.1038/msb.2011.65. PubMed PMID: 21988831; PubMed Central PMCID: PMCPMC3261703.

129. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc. 2007;2(3):727-38. doi: 10.1038/nprot.2007.99. PubMed PMID: 17406635.

130. King ZA, Drager A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. PLoS Comput Biol. 2015;11(8):e1004321. doi: 10.1371/journal.pcbi.1004321. PubMed PMID: 26313928; PubMed Central PMCID: PMCPMC4552468.

131. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO. COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. bioRxiv. 2017. doi: 10.1101/106559.

132. Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. solveME: fast and reliable solution of nonlinear ME models. BMC Bioinformatics. 2016;17(1):391. doi: 10.1186/s12859-016-1240-1. PubMed PMID: 27659412; PubMed Central PMCID: PMCPMC5034503.

133. Ma D, Yang L, Fleming RM, Thiele I, Palsson BO, Saunders MA. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. Sci Rep. 2017;7:40863. doi: 10.1038/srep40863. PubMed PMID: 28098205; PubMed Central PMCID: PMCPMC5241643.

134. Palmer KL, Kos VN, Gilmore MS. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. Curr Opin Microbiol. 2010;13(5):632-9. doi: 10.1016/j.mib.2010.08.004. PubMed PMID: 20837397; PubMed Central PMCID: PMCPMC2955785.

135. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. 2014;5:172. doi: 10.3389/fmicb.2014.00172. PubMed PMID: 24860555; PubMed Central PMCID: PMCPMC4029002.

136. Chubukov V, Mukhopadhyay A, Petzold CJ, Keasling JD, Martin HG. Synthetic and systems biology for microbial production of commodity chemicals. NPJ Syst Biol Appl. 2016;2:16009. doi: 10.1038/npjsba.2016.9. PubMed PMID: 28725470; PubMed Central PMCID: PMCPMC5516863 study design, data collection and analysis, decision to publish, or preparation of the manuscript.

137. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science. 2015;348(6237):921-5. doi: 10.1126/science.aaa0769. PubMed PMID: 25999509; PubMed Central PMCID: PMCPMC4718922.

138. Kachroo AH, Laurent JM, Akhmetov A, Szilagyi-Jones M, McWhite CD, Zhao A, Marcotte EM. Systematic bacterialization of yeast genes identifies a near-universally swappable pathway. Elife. 2017;6. doi: 10.7554/eLife.25093. PubMed PMID: 28661399.

139. Lind PA, Tobin C, Berg OG, Kurland CG, Andersson DI. Compensatory gene amplification restores fitness after inter-species gene replacements. Mol Microbiol. 2010;75(5):1078-89. doi: 10.1111/j.1365-2958.2009.07030.x. PubMed PMID: 20088865.

140. Kacar B, Ge X, Sanyal S, Gaucher EA. Experimental Evolution of Escherichia coli Harboring an Ancient Translation Protein. J Mol Evol. 2017;84(2-3):69-84. doi: 10.1007/s00239-017-9781-0. PubMed PMID: 28233029; PubMed Central PMCID: PMCPMC5371648.

141. Bershtein S, Serohijos AW, Bhattacharyya S, Manhart M, Choi JM, Mu W, Zhou J, Shakhnovich EI. Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. PLoS Genet. 2015;11(10):e1005612. doi: 10.1371/journal.pgen.1005612. PubMed PMID: 26484862; PubMed Central PMCID: PMCPMC4618355.

142. McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in Escherichia coli pgi knockout evolutions. Appl Environ Microbiol. 2018. doi: 10.1128/AEM.00823-18. PubMed PMID: 30054360.

143. Long CP, Gonzalez JE, Feist AM, Palsson BO, Antoniewicz MR. Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in Escherichia coli. Proc Natl Acad Sci U S A. 2018;115(1):222-7. doi: 10.1073/pnas.1716056115. PubMed PMID: 29255023; PubMed Central PMCID: PMCPMC5776819.

144. McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. Adaptation to the coupling of glycolysis to toxic methylglyoxal production in tpiA deletion strains of Escherichia coli requires synchronized and counterintuitive genetic changes. Metabolic engineering. 2018;48:82-93. doi: 10.1016/j.ymben.2018.05.012. PubMed PMID: 29842925.

145. Lenski RE. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. ISME J. 2017;11(10):2181-94. doi: 10.1038/ismej.2017.69. PubMed PMID: 28509909; PubMed Central PMCID: PMCPMC5607360.

146. de Avila ESS, Notari DL, Neis FA, Ribeiro HG, Echeverrigaray S. BacPP: a web-based tool for Gram-negative bacterial promoter prediction. Genet Mol Res. 2016;15(2). doi: 10.4238/gmr.15027973. PubMed PMID: 27173187.

147. Kershner JP, McLoughlin SY, Kim J, Morgenthaler A, Ebmeier CC, Old WM, Copley SD. A Synonymous Mutation Upstream of the Gene Encoding a Weak-Link Enzyme Causes an Ultrasensitive Response in Growth Rate. J Bacteriol. 2016;198(20):2853-63. doi: 10.1128/Jb.00262-16. PubMed PMID: WOS:000384347500012.

148. Peil L, Starosta AL, Lassak J, Atkinson GC, Virumae K, Spitzer M, Tenson T, Jung K, Remme J, Wilson DN. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. Proc Natl Acad Sci U S A. 2013;110(38):15265-70. doi: 10.1073/pnas.1310642110. PubMed PMID: 24003132; PubMed Central PMCID: PMCPMC3780873.

149. Bailey SF, Hinz A, Kassen R. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. Nat Commun. 2014;5:4076. doi: 10.1038/ncomms5076. PubMed PMID: 24912567.

150. Agashe D, Sane M, Phalnikar K, Diwan GD, Habibullah A, Martinez-Gomez NC, Sahasrabuddhe V, Polachek W, Wang J, Chubiz LM, Marx CJ. Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. Molecular biology and evolution. 2016;33(6):1542-53. doi: 10.1093/molbev/msw035. PubMed PMID: 26908584; PubMed Central PMCID: PMCPMC4868122.

151. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA. NUPACK: Analysis and design of nucleic acid systems. J Comput Chem. 2011;32(1):170-3. doi: 10.1002/jcc.21596. PubMed PMID: 20645303.

152. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic Escherichia coli. Nature. 2000;406(6791):64-7. doi: 10.1038/35017546. PubMed PMID: 10894541.

153. Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, Kim D, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Szubin R, Hefner Y, Feist AM, Palsson BO. Cellular responses to reactive oxygen species can be predicted on multiple biological scales from molecular mechanisms. bioRxiv. 2018. doi: 10.1101/227892.

154. Guzman GI, Olson CA, Hefner Y, Phaneuf PV, Catoiu E, Crepaldi LB, Micas LG, Palsson BO, Feist AM. Reframing gene essentiality in terms of adaptive flexibility. BMC Syst Biol. 2018.

155. Liu B, Hong C, Huang RK, Yu Z, Steitz TA. Structural basis of bacterial transcription activation. Science. 2017;358(6365):947-51. doi: 10.1126/science.aao1923. PubMed PMID: 29146813.

156. Wytock TP, Fiebig A, Willett JW, Herrou J, Fergin A, Motter AE, Crosson S. Experimental evolution of diverse Escherichia coli metabolic mutants identifies genetic loci for convergent adaptation of growth rate. PLoS Genet. 2018;14(3):e1007284. doi: 10.1371/journal.pgen.1007284. PubMed PMID: 29584733; PubMed Central PMCID: PMCPMC5892946.

157. Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. Codon Usage Selection Can Bias Estimation of the Fraction of Adaptive Amino Acid Fixations. Molecular biology and evolution. 2016;33(6):1580-9. doi: 10.1093/molbev/msw027. PubMed PMID: 26873577.

158. McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO. Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. Nat Commun. 2018;9(1):3796. doi: 10.1038/s41467-018-06219-9.

159. Herring CD, Glasner JD, Blattner FR. Gene replacement without selection: regulated suppression of amber mutations in Escherichia coli. Gene. 2003;311:153-63. PubMed PMID: 12853150.

160. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular biology and evolution. 2017;34(7):1812-9. doi: 10.1093/molbev/msx116. PubMed PMID: 28387841.

161. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 2015;43(W1):W580-4. doi: 10.1093/nar/gkv279. PubMed PMID: 25845596; PubMed Central PMCID: PMCPMC4489272.

162. Sandberg TE, Lloyd CJ, Palsson BO, Feist AM. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. Appl Environ Microbiol. 2017. doi: 10.1128/AEM.00410-17. PubMed PMID: 28455337.

163. Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. DNA extraction for streamlined metagenomics of diverse environmental samples. Biotechniques. 2017;62(6):290-3. doi: 10.2144/000114559. PubMed PMID: 28625159.

164. Wong BG, Mancuso CP, Kiriakov S, Bashor CJ, Khalil AS. Precise, automated control of conditions for high-throughput growth of yeast and bacteria with eVOLVER. Nature biotechnology. 2018;36(7):614-23. doi: 10.1038/nbt.4151. PubMed PMID: 29889214; PubMed Central PMCID: PMCPMC6035058.

165. Radek A, Tenhaef N, Muller MF, Brusseler C, Wiechert W, Marienhagen J, Polen T, Noack S. Miniaturized and automated adaptive laboratory evolution: Evolving Corynebacterium glutamicum towards an improved d-xylose utilization. Bioresour Technol. 2017;245(Pt B):1377-85. doi: 10.1016/j.biortech.2017.05.055. PubMed PMID: 28552568.

166. Phaneuf PV, Gosting D, Palsson B, Feist A. ALEdb 1.0: A Database of Mutations from Adaptive Laboratory Evolution. bioRxiv. 2018. doi: 10.1101/320747.

167. Din MO, Danino T, Prindle A, Skalak M, Selimkhanov J, Allen K, Julio E, Atolia E, Tsimring LS, Bhatia SN, Hasty J. Synchronized cycles of bacterial lysis for in vivo delivery. Nature. 2016;536(7614):81-5. doi: 10.1038/nature18930. PubMed PMID: 27437587; PubMed Central PMCID: PMCPMC5048415.

168. Thyer R, Shroff R, Klein DR, d'Oelsnitz S, Cotham VC, Byrom M, Brodbelt JS, Ellington AD. Custom selenoprotein production enabled by laboratory evolution of recoded bacterial strains. Nature biotechnology. 2018;36(7):624-31. doi: 10.1038/nbt.4154. PubMed PMID: 29863724; PubMed Central PMCID: PMCPMC6035053.

169. Wannier TM, Kunjapur AM, Rice DP, McDonald MJ, Desai MM, Church GM. Adaptive evolution of genomically recoded Escherichia coli. Proc Natl Acad Sci U S A. 2018;115(12):3090-5. doi: 10.1073/pnas.1715530115. PubMed PMID: 29440500; PubMed Central PMCID: PMCPMC5866557.

170. Halperin SO, Tou CJ, Wong EB, Modavi C, Schaffer DV, Dueber JE. CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. Nature. 2018;560(7717):248-+. doi: 10.1038/s41586-018-0384-8. PubMed PMID: WOS:000441115200051.