2018

Methods for analysis of derivative strains from metabolic evolution experiments

Erin Boggess
Iowa State University

Recommended Citation

Boggess, Erin, "Methods for analysis of derivative strains from metabolic evolution experiments" (2018). *Graduate Theses and Dissertations*. 16766.
https://lib.dr.iastate.edu/etd/16766

Methods for analysis of derivative strains from metabolic evolution experiments

by

Erin E. Boggess

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Julie Dickerson, Major Professor
Laura Jarboe
Alicia Carriquiry
David Oliver
Zengyi Shao

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

Page

LIST OF FIGURES v

LIST OF TABLES vii

ACKNOWLEDGMENTS viii

ABSTRACT ix

CHAPTER 1. INTRODUCTION 1
Metabolic Evolution in Bacteria 1
Octanoic acid toxicity in *E. coli* 3
Octanoic acid-tolerant *E. coli* 5
Genomic Mutations 5
Genotype-Phenotype Relationship 7
Goal of this work 8
Thesis Organization 10
References 12

CHAPTER 2. MUTATION ANALYSIS FOR METABOLIC EVOLUTION EXPERIMENTS IN *ESCHERICHIA COLI* 15
Abstract 15
Background 15
Results 15
Conclusion 16
Background 16
Methods 22
Mutation Annotation 22
Analysis and Prioritization 24
Genes 24
Structural Features and RNAs 27
Binding Sites 28
Gene regulatory and metabolic network generation 30
Software implementation 33

Results 34
Antibiotic resistance in *E. coli* 34
Cra regulon, *marRAB* operon, and multidrug efflux system 39
LexA regulon 40
Benzoate tolerance in *E. coli* 41
Discussion 46
Conclusions 48
Acknowledgements 48

Authors' contributions ... 49
References ... 49

CHAPTER 3. GENOME-LEVEL REVERSE ENGINEERING OF *ESCHERICHIA COLI EVOLVED* FOR INCREASED SHORT-CHAIN FATTY ACID TOLERANCE AND PRODUCTION ... 56
Abstract ... 56
Introduction ... 57
Materials and methods ... 60
Strains, plasmids and bacterial cultivation ... 60
Whole-genome sequencing and mutation verification ... 63
PCR-restriction fragment length polymorphism (PCR-FRLP) ... 64
Genomic manipulations ... 65
Octanoic acid tolerance test ... 65
Determination of fatty acid titers ... 66
Membrane characterization ... 67
Membrane permeability ... 67
Membrane fluidity ... 67
Cell surface hydrophobicity ... 68
Membrane lipid composition ... 68
Results and Discussion ... 69
Verified mutations in evolved strains ... 69
Colony morphology of parent and evolved strains ... 75
The order in which mutations were acquired during the metabolic evolution experiment ... 75
Growth ability of reconstructed strains in C8 challenge experiments ... 76
The *waaG*, *basS* and *basR* mutations affect the cell membrane ... 80
The *waaG* and *rpoC* mutations improve fatty acid titer ... 90
Summary ... 91
Acknowledgements ... 93
Authors' Contributions ... 93
References ... 93

CHAPTER 4. TRANSCRIPTOMIC ANALYSIS OF *ESCHERICHIA COLI* EVOLVED FOR OCTANOIC ACID TOLERANCE ... 99
Abstract ... 99
Introduction ... 99
Methods ... 101
Bacterial strains, plasmids, and media ... 101
Growth analysis ... 102
RNA Isolation ... 103
Fermentation for fatty acid production ... 103
RNA-seq and short read analysis ... 104
Results and discussion ... 106
Fatty acid fermentation for ML115 and LAR1 with pJMY-EEI82564 or pJMY-Empty ... 106

Identifying the *rpoC* mutation effect among differentially expressed genes 107
Expression of genes in the BasR regulon 112
bssS knockout improves specific growth rate at 10 mM C8 113
Conclusions 119
References 119

CHAPTER 5. SUMMARY, FUTURE WORK, AND CONCLUSIONS 123

APPENDIX A. IDENTIFICATION OF MUTATIONS IN EVOLVED BACTERIAL GENOMES 126
Summary 126
Introduction 126
Materials 128
Genomic DNA Purification and Sequencing 129
Bioinformatics Software for High-Throughput Sequence Data 129
Mutation Verification 131
Methods 132
Obtain Sequence Data 132
Preprocess Sequence Data 134
Map Short Reads to Reference Genome 137
De novo Assembly 138
Identify Variations in Evolved Strains 139
Verify Mutations 140
Notes 143
Acknowledgements 146
Refereences 146
Tables 150
Figures 152

APPENDIX B. *E. COLI* VARIANT ANALYSIS (EVA) USER GUIDE 155
Software Requirements 155
Commands 155
Configuration 158

Hierarchy for all packages 161
Class Hierarchies 161
Interface Hierarchy 166
Enum Hierarchy 166

LIST OF FIGURES

Page

Figure 1.1 Microbial engineering to achieve a desired phenotype. 3

Figure 1.2 Overview of EVA software design. 9

Figure 2.1 EVA pipeline 21

Figure 2.2 Gene model 24

Figure 2.3 Combined gene regulatory and metabolic network structure 31

Figure 2.4 Network visualization of mutations in antibiotic resistant strains 38

Figure 2.5 Network visualization of mutations in benzoate-adapted strains.................. 45

Figure 3.1 EVA-generated network. 73

Figure 3.2 An insertion sequence interrupts *waaG* in ML115. 75

Figure 3.3 Morphological characteristic of ML115 and LAR1.................................... 77

Figure 3.4 Transmission electron microcopy images of ML115 and LAR1.................. 77

Figure 3.5 PCR experiments to determine the order of mutations. 78

Figure 3.6 Extracellular polymetric substance analysis for ML115, LAR1, and reconstructed strains. 78

Figure 3.7 Growth ability of ML115, LAR1, and reconstructed strains in the order that mutations were acquired. 81

Figure 3.8 Membrane polarization.. 86

Figure 3.9 Membrane leakage 87

Figure 3.10 Percent of hydrophobicity... 88

Figure 3.11 Membrane lipid composition. 89

Figure 3.12 Fatty acid titer measured over time. 91

Figure 4.1 Strain characteristics during fatty acid fermentation................................. 108

Figure 4.2 Fatty acid beta oxidation pathway.. 111

Figure 4.3 EVA-generated gene regulatory and metabolic networks for genes in Table 3.2. 114

Figure 4.4 EVA-generated network with RNA-seq data. 117

Figure 4.5 The effect on specific growth rate of Δ*bssS* in 0 and 10 mM C8. 118

LIST OF TABLES

Page

Table 2.1 Prioritization of mutations in protein-coding genes 25

Table 2.2 A selection of previously unreported predicted mutations in antibiotic-resistant strains 36

Table 2.3 Summary of nodes and edges in biological networks generated by EVA....... 37

Table 2.4 Previously reported intergenic mutations in one or more benzoate-tolerant strains for which EVA provides annotations. 43

Table 3.1 Plasmids used in this study. 61

Table 3.2 Strains used in this study.. 62

Table 3.3 EVA annotation and prioritization of mutations in ML115, LAR1, and LAR2. 72

Table 3.4 Growth ability of ML115, LAR1, and reconstructed strains. 82

Table 4.1 Strains and plasmids used in this study. 102

Table 4.2 Genes with consistent and statistically significant fold changes for all strain contrasts.. 115

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Julie Dickerson, and my committee members, Dr. Laura Jarboe, Dr. Zengyi Shao, Dr. David Oliver, and Dr. Alicia Carriquiry, for their guidance and support throughout the course of this research. I would also like to thank Dr. Jackie Shanks for her collaboration and encouragement. Additionally, I want to thank members of the Dickerson lab and the NSF Engineering Research Center for Biorenewable Chemicals (CBiRC), both past and present, who have provided invaluable advice and feedback and contributed to my success over the years. In particular, I would like to acknowledge my research collaborators in CBiRC, Liam Royce and Yingxi Chen for their hard work and dedication.

This work was funded primarily through a fellowship made possible by the National Science Foundation grant EEC-0813570. Additional funding was made available through a fellowship award under the Integrative Graduate Education and Research Training (IGERT) grants from the National Science Foundation.

ABSTRACT

One of the largest challenges in genomics studies is determining the relationship between genotype and phenotype and then applying this knowledge to design principles. Metabolic engineering of bacteria can introduce targeted genomic interventions to well-characterized genes for the purpose of modifying cellular metabolism, but in some cases, even for the model organism *Escherichia coli*, alternative strategies are required to achieve a desired phenotype. Metabolic evolution involves applying selective pressure to a population, and over time advantageous mutations will arise that improve organism fitness. To understand what mutations occurred during these experiments and how they affect phenotype, whole genome sequencing is required, followed by mutation analysis and strain characterization.

Genome sequencing generates a large amount of data for researchers to examine and traditionally mutation analysis focuses only on gene variations. Supporting mutation analysis with computational tools and using a systems-level approach that utilizes public databases describing gene regulation and cellular metabolism improves upon existing analysis techniques and advances our understanding of how genotype relates to phenotype.

Using our mutation analysis software, *E. coli* Variant Analysis (EVA), we examine antibiotic resistance, benzoate tolerance, and octanoic acid tolerance in *E. coli*. Our analysis pipeline includes a defined set of rules for mutation categorization. Prioritization of mutations supports efforts to reverse-engineer evolved strains and focus on the variants most likely to be damaging or relevant to phenotype. From mutation analysis results, we construct biological networks for visualization of mutations and possible downstream effects. This allows for improved mutation interpretation and identification of possible mutation interactions. Furthermore, we integrate RNA-seq data into our analysis to investigate the effects of variant regulators on the transcriptome. In contrast to existing methods which focus on mutated genes, we incorporate annotations for binding sites and other regulatory features on the genome for the most complete interpretation based on the available genome and gene regulatory models.

CHAPTER 1. INTRODUCTION

Metabolic Evolution in Bacteria

Microbial metabolic engineering is an essential tool for generating organisms capable of specialized chemical production. The Center for Biorenewable Chemicals aims to use a combination of biocatalysis and chemical catalysis methods in order to establish a sustainable system to produce bio-based chemicals using a model organism such as *Escherichia coli* and *Saccharomyces cerevisiae*. The biocatalysis component involves engineering microbial strains by adding extrinsic functionalities, such as introducing new enzymes, and redirecting target metabolic pathways. In such a way, it is possible to develop a biological system that uses carbohydrate feedstock as input for the production of target chemicals.

Rational design introduces genome modifications in a purposeful manner based on knowledge of gene models and metabolic pathways. Gene knockouts and gene overexpression can modify metabolic pathways to direct carbon flow toward a desired process. Incorporation of novel enzymes can add new synthesis pathways to an organism for production of desired chemicals. However, target chemicals at high titers can cause toxicity which must be overcome for economically feasible production.

Without fully characterizing the toxicity effect, metabolic evolution can be employed to generate a strain that exhibits improved tolerance. Metabolic evolution is a powerful method that requires minimal knowledge of the platform organism or underlying biological mechanisms to obtain a strain with a desired phenotype. It is considered a black box technique because scientists do not control or observe genetic changes as they occur. In a metabolic evolution experiment, variant strains with advantageous phenotypes emerge under strong selective pressure and displace the parent strain in a population. Improved fitness is attributed to one or more mutations acquired during the experiment. Metabolic evolution may be repeated along with additional genomic interventions to obtain a desired phenotype. Reverse engineering of the evolved strain is required to elucidate which mutations are relevant to fitness and by what mechanisms. Reverse engineering involves a comparative analysis of the genomes of the parental strain and evolved strain to identify mutations and characterization of each mutation as well as their combined effects. In addition to comparative genomic analysis, additional omics studies may be performed to characterize evolved strain(s) and identify factors that contribute to improved fitness. Reverse engineering also requires reconstructing verified mutations in the parent strain and repairing mutations in evolved strains for comparative analyses. An overview of this design strategy for strain design is provided in Figure 1.1. Bioinformatics can aid in reverse engineering by predicting which variations may be damaging to genomic features and how mutated elements may affect other biological features through regulation and cellular metabolism. Understanding how sequence variations can affect genes, binding sites, and other genomic sequences is critical to characterizing an evolved strain. In some cases, biological elements with regulatory features are mutated, which have the potential to lead to large scale changes in gene expression. Gene regulatory network data can help identify downstream elements which could be indirectly affected such a mutation. Integrating this information with pathway data can highlight metabolic activities that may be altered and have relevance to phenotype.

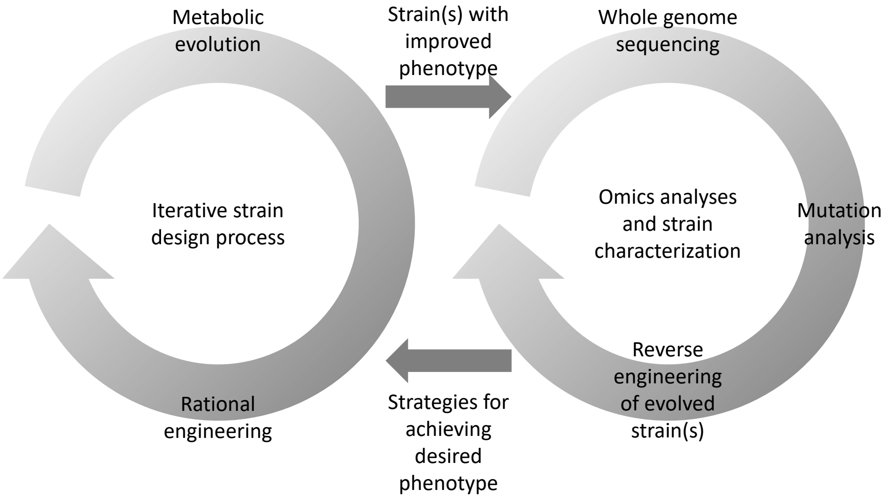

Metabolic evolution
Strain(s) with improved phenotype
Whole genome sequencing
Iterative strain design process
Omics analyses and strain characterization
Mutation analysis
Reverse engineering of evolved strain(s)
Rational engineering
Strategies for achieving desired phenotype

Figure 1.1 *Microbial engineering to achieve a desired phenotype.*

Both metabolic evolution and rational engineering strategies may be used in an iterative manner to achieve a desired phenotype. Metabolic evolution requires reverse engineering of evolved strain(s) to identify key mutations and understand their relevance to phenotype.

Octanoic acid toxicity in *E. coli*

Carboxylic acid is a testbed in *E. coli* from which CBiRC envisions generating multiple biorenewable chemical products. Fatty acids are useful in industrial applications including surfactants and lubricants (Makkar & Cameotra, 2002) and are an attractive target chemical due to the ability to break the elongation cycle at varying chain lengths by introducing foreign acyl-ACP thioesterases with different specificities (X. Zhang, Li, Agrawal, & San, 2011). However, at high concentrations, fatty acids become toxic to the *E. coli*, limiting titers, yields, and productivity. Specifically, fatty acid toxicity in *E. coli* has been shown to be related to membrane damage (Desbois & Smith, 2010; R. M. Lennen et al., 2011). Toxicity effects on the membrane were also investigated in an octanoic acid (C8) challenge by characterizing the effect on the membrane (L. A. Royce, P. Liu, M. J. Stebbins,

B. C. Hanson, & L. R. Jarboe, 2013), confirming that short chain fatty acids damage cell membrane by increasing membrane fluidity and porosity.

To further characterize the mechanisms of fatty acid toxicity, we performed transcriptome analysis of *E. coli* during exogenous challenge of octanoic acid (C8) (Royce et al., 2014). In the experiment, *E. coli* K-12 MG1655 was grown to mid-log phase in MOPS minimal media with and without 10 mM C8 at initial pH of 7.0. RNA was hybridized to AffyMetrix GeneChip *E. coli* Genome 2.0 Arrays for three biological replicates of each condition and analyzed at the DNA facility of Iowa State University. I performed background adjustment, normalization, and summarization calculations in MATLAB using GCRNA (Irizarry, Wu, & Jaffee, 2006; Z. Wu, Irizarry, Gentleman, Martinez-Murillo, & Spencer, 2004). Many genes with increased expression in response to the C8 challenge were related to acid response, response to and regulation of pH, and biofilm formation. Genes with decreased expression were related to reduced motility, chemotaxis, and flagellum assembly. In addition, we identified perturbed genes associated with membrane function and integrity: *bhsA*, *cpxP*, *cfa*, and *ompF*. I also performed Network Component Analysis (NCA) (Fu, Jarboe, & Dickerson, 2011; Liao et al., 2003) to predicted altered transcription factor activity based on changes in transcript abundance of regulated genes. Notably, the transcription factor GadE (glutamate-dependent acid resistance system) was predicted to have altered activity, attributed to increased expression in acid resistance genes *hdeABD* and *gadABCE* in response to C8. Due to the significant increase in activity of the GadE regulon, glutamate supplementation was tested as a method to increase tolerance C8 but did not prove effective. We hypothesize that membrane damage impairs the glutamate-dependent acid resistance system during C8 challenge.

Octanoic acid-tolerant *E. coli*

To overcome toxicity limitations, a more robust strain was engineered using a combination of rational design and metabolic evolution methods. The parent strain for the metabolic evolution experiment was ML115, which had been engineered with three deletions (*fadD*, *ack-pta*, and *poxB*) to inactivate the fatty acid beta-oxidation pathway to halt fatty acid degradation and two acetate pathways to redirect flux to increase the acetyl-CoA pool (M. Li, Zhang, Agrawal, & San, 2012). Microbial metabolic evolution was performed by 15 sequential transfers with increasing C8 concentration from 10 mM to 30 mM over 714 hours. At the end of the experiment, evolved strains LAR1 and LAR2 were isolated for genome sequencing along with the parent strain (L. A. Royce et al., 2015).

Genomic DNA of evolved strains LAR1, LAR2, and the parent strain, ML115, was isolated and sequenced using Illumina whole genome sequencing. I aligned short reads to the MG1655 reference genome and predicted mutations using methods outlined in (L. Royce, E. Boggess, T. Jin, J. Dickerson, & L. Jarboe, 2013) and found in Appendix A. Mutations predicted in both the ancestral and evolved strains were not considered for further analysis. A key mutation that was found in both evolved strains occurred in *rpoC*, which encodes the β' subunit of RNA polymerase. A point mutation from A to C at position 1256 in *rpoC* results in an amino acid change of H419P. Two other mutations were predicted in *basR* (LAR1) and *basS* (LAR2) genes. In *basR*, a point mutation of G to T at position 82 results in an amino acid change of D28Y. In *basS*, a 27 base pair (bp) deletion results in the deletion of nine amino acids in the protein BasS.

Genomic Mutations

Spontaneous mutations may result from errors in DNA replication, DNA lesions, and transposable elements. Errors in DNA replication can arise when mis-paired nucleotides result in a base substitution and strand slips at repeated sequences. Naturally occurring damage to DNA can cause spontaneous lesions that can lead to mutations. Depurination, the spontaneous loss of the glycosidic bond between a base and deoxyribose can lead to a substitution or loss of a nucleotide pair. Deamination of cytosine to uracil results in a base substitution. Oxidative damage can also cause DNA lesions leading to mutagenesis. In rare cases, the DNA molecule itself may break and in the act of non-homologous end joining nucleotides may be added or removed to repair the molecule (Moore & Haber, 1996). Finally, transposable elements, or "jumping genes," consist of DNA sequences that are capable of moving and inserting into the genome at new positions. In a laboratory setting, increased mutation rates can be achieved with the use of chemicals that destabilize DNA molecules and by irradiation (e.g., ultraviolet light) (Lee, Feist, Barrett, & Palsson, 2011).

Genomic mutations in bacterial systems can be categorized by their impact on DNA sequence. The following mutation types apply to bacterial systems in evolution studies: substitution, deletion, insertion, indel, amplification, and translocation. Substitution mutation describe replacing bases with an alternative sequence. A single nucleotide polymorphism (SNP) describes the substitution of a single base ((A)denine, (C)ytosine, (G)uanine, (T)hymine) for another. Deletions and insertions describe the excision or addition of nucleotides in the genome, respectively. In some cases, a more complex mutation may occur that results in a deletion and insertion of unequal lengths, which is known as an indel (insertion-deletion of DNA). Amplifications describe the replication of a DNA sequence and translocations refer to a DNA sequence that is relocated to a different position on the genome.

Furthermore, mutations can be described by their change to protein-coding genes. Silent mutations describe an altered DNA sequence, but no change in the amino acid sequence. Missense mutations describe the substitution of an amino acid for another. A special case is the substitution of an amino acid for a premature stop codon, which is known as a nonsense mutation. A mutation that results in the loss of a stop codon is known as a nonstop or read-through mutation and transcription may continue until the next stop codon is encountered. In some cases, stop and start codons are preserved, but a DNA mutation results in an alternate stop/start codon that can alter translation initiation rates (Hecht et al., 2017) or require alternative release factors (Korkmaz, Holm, Wiens, & Sanyal, 2014). In-frame insertions, deletions, an indels that result in the insertion or deletion of amino acids and out-of-frame mutations cause frameshifts that modify downstream codons and may result in loss of function.

Genotype-Phenotype Relationship

As the price of DNA sequencing continues to fall, more genomic data will continue to be generated for metabolic evolution and comparative studies. However, using the genetic sequence of an organism to predict its phenotype is an open biological problem. A similar ambition is to measure features of organism phenotype through molecular and cellular experiments and trace these characteristics back to features on the genome. An understanding the relationship between the genome and phenotypic traits aids both goals. Even for model organisms, discerning genotype-phenotype relationships remains a challenge as our knowledge of biological systems is incomplete and existing models are composed of entangled networks of regulatory activities such that altering one element may affect many other features.

Mutations in coding regions can alter transcript abundance, product abundance, and product function. Some genes encode transcription factors which regulate transcription of other genes. A variant transcription factor may show altered activity or binding site specificity, resulting in altered expression of regulated genes. In addition to genes, mutations in regulatory sites can contribute to the phenotype of evolved strains. Regulation of gene expression is a critical response mechanism to environmental stimuli and integral to controlling cellular behavior. Altering parameters such binding site affinity, transcription factor abundance, and regulatory elements functioning as secondary structures can also affect transcription and translation regulation. Gene products either directly (via genomic mutations) or indirectly (via regulation) changed by mutations can affect cellular structure or metabolism resulting in an observed phenotypic trait. Additionally, multiple mutations and may be additive, synergistic, or antagonistic in nature (Elena & Lenski, 1997; Szathmáry, 1993).

Goal of this work

After a metabolic engineering experiment, a bottleneck occurs when mapping genetic modifications to phenotype. Analysis of strains obtained through metabolic evolution traditionally involves manual annotation of mutations in coding regions and evaluation of their individual contribution to fitness through functional or comparative studies. Typically, analysis does not consider extragenic variations (mutations outside of coding regions). The massive amount of sequence variation data generated in evolution experiments necessitates computational tools that can assess mutation implications.

The goal of this project is to construct a framework to systematically analyze mutations and provide interpretations for both direct impact of mutations and potential downstream effects that occur through regulation. In doing so, we aim to support efforts to reverse engineer adapted strains generated from metabolic engineering experiments and reduce the amount of time to a secondary round of metabolic engineering.

To achieve this goal, genes, products, regulatory elements, metabolic pathway information, and relationships of these entities are included in the mutation analysis pipeline. Data is queried from publicly available databases RegulonDB (S. Gama-Castro et al., 2016) and EcoCyc (I. M. Keseler et al., 2017).

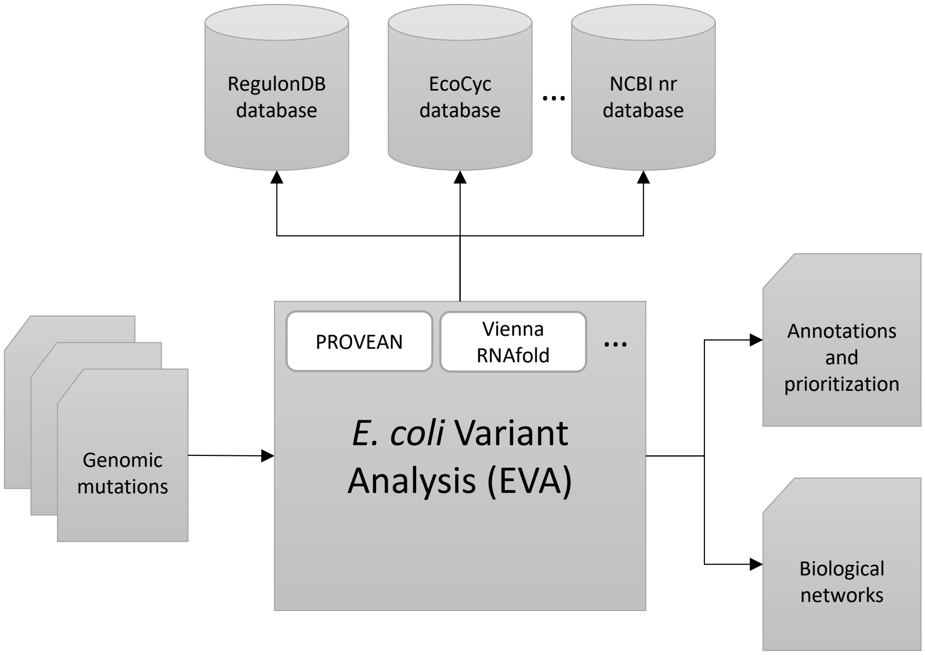


Figure 1.2 *Overview of EVA software design.*

EVA accepts text files that contain positional and sequence information about genomic mutations as input. Annotations for each mutation are obtained by querying publicly available E. coli databases. Depending on the type of mutation, various strategies can be employed for additional analysis, some of which use published protein sequences obtained from NCBI. Gene regulatory and metabolic data are retrieved from publicly available databases and used to construct biological networks that aid in visualizing mutations, their effects, and potential interactions. A text file containing annotations, reference and alternate feature sequences, and other analysis results is also generated as part of EVA's output.

Translating between DNA, RNA, and amino acid molecules and defining a set of rules for types of variations and regulatory activities aid interpretations. Additionally, mutations are examined for their effect on molecular structures and binding affinities to predict if they have a significant impact to the organism. The proposed methods benefit the research community by broadening the study of mutations and mechanisms of adaptation. Additionally, automating portions of comparative genomic analysis reduces the lifecycle of adaptive evolution studies.

Thesis Organization

The following chapters are a collection of research papers and book chapters that are either published, under review, or intended for submission for publication when complete. They are organized as follows: Chapters 2-4 are research papers presented on the topic of mutation analysis for evolved strains in metabolic engineering experiments and elucidating genotype-phenotype relationships. Chapter 5 is a general discussion on the significance and impact of studies presented in Chapters 2-4. Appendix A is a methods chapter on reverse engineering of evolved strains. Appendix B is a user manual for our mutation analysis software. Appendix C contains supplementary material.

Chapter 2: Mutation Analysis for Metabolic Experiments in *Escherichia coli*

This research paper describes *E. coli* Variant Analysis (EVA) software for mutation analysis in *E. coli*. Methods for annotating and interpreting mutations as well as integration with gene regulatory and metabolic networks are presented to investigate mutation effects and aid in elucidating their genotype-phenotype relationship. Additionally, algorithms for network reduction can highlight potential mutation-mutation interactions.

Chapter 3: Genome-level Reverse Engineering of *Escherichia coli* Evolved for Increased Short-Chain Fatty Acid Tolerance and Production

Chapter 3 is a research paper that discusses metabolic engineering as a method for increasing tolerance and production of octanoic acid and genomic analysis of evolved *E. coli* strains. Assembly and analysis of short read sequence data is an integral part of this work. Comparative analysis of evolved and ancestral genomes is required for reverse engineering. Mutation analysis and interpretation identifies potentially damaging variants in the evolved strain, including a global regulator and transcription factor which may alter expression of many regulated genes.

Chapter 4: Transcriptomic Analysis of *Escherichia coli* Evolved for Increased Short-Chain Fatty Acid Tolerance and Production

Chapter 4 is a research paper extends previous work to reverse engineer *E. coli* evolved for octanoic acid tolerance with the addition of RNA-seq experiments. Transcriptomic analysis was performed for an evolved strain (LAR1) and ancestral strain (ML115) in both control and fatty acid production conditions at three time points. Normalization and differential expression analysis led to the identification of genes that were perturbed for all strain contrasts. This list of genes was then annotated with associated sigma factors and presence in the BasR. Genes with significant fold changes were submitted as candidates for further investigation into the effect of previously identified mutations in the global regulator, RpoC, and transcription factor, BasR. The use of EVA software in combination with transcriptomic data was a key component in predicting the effects of previously identified genomic mutations in transcription regulators. EVA was also used to identify relationships among differentially expressed genes and highlight metabolic pathways in which they participate.

Chapter 5: Conclusions and future work

The final chapter summarizes important findings and discusses the significance of the work presented in chapters 2-4. The future work describes recommendations for extending and improving upon the work presented in this dissertation.

Appendix A: Identification of Mutations in Evolved Bacterial Genomes

Reverse engineering of microbial strains evolved in metabolic evolution experiments is necessary to understand the mechanisms that result in a desired phenotype. This book chapter details methods for short read analysis of genomic data and mutation identification.

Appendix B: *E. coli* Variant Analysis (EVA) User Guide

This section is a user guide for EVA software. Various software options and usage are described in detail.

References

Desbois, A. P., & Smith, V. J. (2010). Antibacterial free fatty acids: activities, mechanisms of action and biotechnological potential. *Appl Microbiol Biotechnol, 85*(6), 1629-1642. doi:10.1007/s00253-009-2355-3

Elena, S. F., & Lenski, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature, 390*(6658), 395-398. doi:10.1038/37108

Fu, Y., Jarboe, L. R., & Dickerson, J. A. (2011). Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics, 12*, 233. doi:10.1186/1471-2105-12-233

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res, 44*(D1), D133-143. doi:10.1093/nar/gkv1156

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., . . . Salit, M. (2017). Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res, 45*(7), 3615-3626. doi:10.1093/nar/gkx070

Irizarry, R. A., Wu, Z., & Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics, 22*(7), 789-794. doi:10.1093/bioinformatics/btk046

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., . . . Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res, 45*(D1), D543-D550. doi:10.1093/nar/gkw1003

Korkmaz, G., Holm, M., Wiens, T., & Sanyal, S. (2014). Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem, 289*(44), 30334-30342. doi:10.1074/jbc.M114.606632

Lee, D. H., Feist, A. M., Barrett, C. L., & Palsson, B. (2011). Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli. *PLoS One, 6*(10), e26172. doi:10.1371/journal.pone.0026172

Lennen, R. M., Kruziki, M. A., Kumar, K., Zinkel, R. A., Burnum, K. E., Lipton, M. S., . . . Pfleger, B. F. (2011). Membrane stresses induced by overproduction of free fatty acids in Escherichia coli. *Appl Environ Microbiol, 77*(22), 8114-8128. doi:10.1128/AEM.05421-11

Li, M., Zhang, X., Agrawal, A., & San, K. Y. (2012). Effect of acetate formation pathway and long chain fatty acid CoA-ligase on the free fatty acid production in E. coli expressing acy-ACP thioesterase from Ricinus communis. *Metab Eng, 14*(4), 380-387. doi:10.1016/j.ymben.2012.03.007

Liao, J. C., Boscolo, R., Yang, Y. L., Tran, L. M., Sabatti, C., & Roychowdhury, V. P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A, 100*(26), 15522-15527. doi:10.1073/pnas.2136632100

Makkar, R. S., & Cameotra, S. S. (2002). An update on the use of unconventional substrates for biosurfactant production and their new applications. *Appl Microbiol Biotechnol, 58*(4), 428-434. doi:10.1007/s00253-001-0924-1

Moore, J. K., & Haber, J. E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Molecular and Cellular Biology, 16*(5), 2164-2173. doi:papers3://publication/doi/10.1128/MCB.16.5.2164

Royce, L., Boggess, E., Jin, T., Dickerson, J., & Jarboe, L. (2013). Identification of mutations in evolved bacterial genomes. *Methods Mol Biol, 985*, 249-267. doi:10.1007/978-1-62703-299-5_13

Royce, L. A., Boggess, E., Fu, Y., Liu, P., Shanks, J. V., Dickerson, J., & Jarboe, L. R. (2014). Transcriptomic analysis of carboxylic acid challenge in Escherichia coli: beyond membrane damage. *PLoS One, 9*(2), e89580. doi:10.1371/journal.pone.0089580

Royce, L. A., Liu, P., Stebbins, M. J., Hanson, B. C., & Jarboe, L. R. (2013). The damaging effects of short chain fatty acids on Escherichia coli membranes. *Appl Microbiol Biotechnol, 97*(18), 8317-8327. doi:10.1007/s00253-013-5113-5

Royce, L. A., Yoon, J. M., Chen, Y., Rickenbach, E., Shanks, J. V., & Jarboe, L. R. (2015). Evolution for exogenous octanoic acid tolerance improves carboxylic acid production and membrane integrity. *Metab Eng, 29*, 180-188. doi:10.1016/j.ymben.2015.03.014

Szathmáry, E. (1993). Do deleterious mutations act synergistically? Metabolic control theory provides a partial answer. *Genetics, 133*(1), 127-132.

Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., & Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association 99*(468), 909-917

Zhang, X., Li, M., Agrawal, A., & San, K. Y. (2011). Efficient free fatty acid production in Escherichia coli using plant acyl-ACP thioesterases. *Metab Eng, 13*(6), 713-722. doi:10.1016/j.ymben.2011.09.007

CHAPTER 2. MUTATION ANALYSIS FOR METABOLIC EVOLUTION EXPERIMENTS IN *ESCHERICHIA COLI*

A research paper submitted to BMC Bioinformatics

Erin E. Boggess[1], Laura R. Jarboe[2], Julie A. Dickerson[1]

[1]*Department of Electrical & Computer Engineering, Iowa State University, Ames, IA, 50011.*

[2]*Department of Chemical & Biological Engineering, Iowa State University, Ames, IA, 50011.*

Abstract

Background

Metabolic evolution, a tool used in strain engineering, involves applying selective pressure to induce advantageous mutations to a strain for manipulating characteristics such as tolerance, product yield, and growth properties. This method does not reveal which acquired mutations led to improved fitness, by what mechanism, or how mutated genomic features interact to produce a phenotype.

Results

This work establishes a pipeline for mutation analysis in *Escherichia coli* called *E. coli* Variant Analysis (EVA) that integrates public databases and multiple sequence analysis tools. EVA annotates mutations, applies analysis strategies to predict effects of variations, and constructs a biological network of mutated genomic features and downstream gene regulatory and metabolic pathway features. Biological networks produced by EVA aid in reverse engineering evolved strains. When applied to data from *E. coli* evolution experiments, EVA annotates mutations in non-coding regions that traditional analyses

overlook. Networks generated by EVA visualize regulons downstream of mutated features, mutation interactions, and mechanisms related to enhanced fitness.

Conclusion

EVA advances mutation interpretation by annotating regulatory features and incorporating gene regulatory, signaling pathway, and metabolic pathway data downstream of mutated features. EVA generates biological networks comprised of these features and their interactions to support reverse engineering of evolved strains. Software automation reduces the burden of annotation, interpretation, and prioritization of results, thereby decreasing the time to follow-up experimentation and further rounds of metabolic evolution.

Background

Microbial metabolic engineering can develop specialized strains that exhibit desired phenotypes. For well-characterized organisms, a rational design approach may be used for strain development (Jang et al., 2012). Rational design entails performing targeted genome modifications based on literature evidence, metabolic pathway knowledge, and computational predictions intended to alter enzyme abundance and/or function. When genome changes that enable the desired phenotype are unknown, researchers can perform a metabolic evolution. Under selective pressure, variants with advantageous mutations displace the parent strain in a population. Reverse engineering of evolved strains identifies beneficial genetic variations (Jin, Chen, & Jarboe). Metabolic evolution is often considered a black box technique because scientists do not control or observe genetic changes as they occur.

As next-generation sequencing methods become more accessible and affordable, comparing the entire genomes of the parent strain and one or more evolved derivatives is becoming common in evolutionary experiments (Barrick et al.; Haft et al.; Harden et al.; Herring et al.; LaCroix et al.). In these experiments, researchers align short-read data to the parent genome or a published wild-type ancestral genome and predict genetic variations between the reference and aligned strains with SNP-calling algorithms. Improved fitness in an evolved strain is attributed to genetic variations not present in the parent strain. Determining which mutations are random and which contribute to the evolved phenotype remains a reverse engineering challenge that requires a considerable amount of further research and experimentation.

Traditionally, researchers inspect genomic variations in coding regions to determine if they disrupt gene transcription or protein function and search the literature for relevance to the observed phenotype (Byrne et al.; Utrilla et al.). While this can reveal important changes such as loss of function, it ignores mutations in non-coding regions responsible for regulatory activities. In other studies, researchers examine every predicted mutation, but this assumes all have equal importance and requires researchers to construct numerous variant strains to carry out necessary follow-up studies (Atsumi et al.). Parallel evolution experiments for multiple populations can identify commonalities in independently acquired mutations (Sandberg et al.). A more complex experimental design, VERT (Visualizing Evolution in Real-Time), involves collecting intermediate samples at various time points during the evolution experiment for populations in competition (Reyes, Winkler, & Kao, 2012). The VERT method provides insight into the order of acquired mutations and their relation to organism fitness. Regardless of experimental design, mutation interpretation remains a challenge.

Methods to predict if amino acid sequence variants are tolerated or damaging include SDM (Site Directed Mutator) (Worth, Preissner, & Blundell), SIFT (Sorting Intolerant From

Tolerant) (Ng & Henikoff), and Provean (Protein Variance Effect Analyzer) (Choi, Sims, Murphy, Miller, & Chan). Each algorithm uses residue conservation statistics and amino acid properties to predict the likelihood of observing an amino acid substitution. These computational methods classify amino acid sequence variations into two general cases: tolerated and damaging. Tolerated variations are predicted not to affect protein function and damaging variations are predicted to impair protein function. Computational methods capable of predicting the introduction of novel functions due to a change in the amino acid sequence do not currently exist.

The SDM algorithm relies upon amino acid substitution frequencies for families of homologous proteins with available structures and requires a Protein Data Bank (PDB) as input. SDM predicts disruptive mutations based on a stability score that describes the change in free energy between the wild-type and predicted variant protein structures and conservation of structural features. SDM offers a unique mutation analysis, however many proteins lack a PDB structure and proteins with more than one structure require researchers to choose which is most relevant to their experimental conditions, a constraint that is not readily automated.

SIFT predicts the impact of missense mutations based on residue conservation calculated from BLAST multiple sequence alignments of homologous sequences and amino acid properties. The software accepts amino acid variations and their relative positions in a protein but does not support nucleotide variations in the genome as input. SIFT assumes that variants which occur naturally in highly similar sequences are less likely to disrupt a protein than variants that are rarely or not at all observed. SIFT may accept National Center for

Biotechnology Information (NCBI) protein IDs or an amino acid sequence in FASTA format as input.

Provean is similar to SIFT, but employs the added step of clustering BLAST results and uses the BLOSUM62 (Henikoff & Henikoff) matrix for scoring amino acid substitutions rather than constructing a PSSM matrix based on BLAST results. Multiple sequence alignments are performed between the query amino acid sequence and clusters of related sequences and alignment scores are averaged. An amino acid variation is predicted to be damaging if the value is below a threshold (authors recommend a cutoff of $\leq$ -2.5 based on testing performed with the UniProt human dataset). Like SIFT, Provean is primarily intended for human studies but can be used for other organisms by submitting the wild-type amino acid sequence and a description of the variation. The description format follows Human Genome Variant Society (HGVS) format (J. T. den Dunnen & Antonarakis; Johan T. den Dunnen et al.).

In addition to genes, mutations in regulatory sites can contribute to the phenotype of evolved strains. Regulation of gene expression is a critical response mechanism to environmental stimuli and integral to controlling cellular behavior. Tuning parameters such as RNA polymerase (RNAP) binding affinity, transcription factor abundance, transcription factor binding site (TFBS) affinity, and ribosome binding site (RBS) affinity can control RNA and protein abundance. One of the most extensively studied regulatory sites is the σ^{70} promoter (Finn; Malhotra, Severinova, & Darst). It has been shown that point mutations in the consensus sequence can result in a broad range of gene expression control (Bakke et al.). For promoters and other binding sites, high-throughput studies on a large number of sequence variations to determine affinity and specificity (Stormo & Zhao). Changes to non-coding RNAs and regulatory elements functioning as secondary structures can also affect transcription and translation (Yu, Bing, & Zhenhua) and several metrics exist to describe conformational differences between wild-type and variant sequences (Avihoo & Barash, 2006; Kiryu & Asai; Sabarinathan et al.).

The proposed EVA workflow includes the following components: annotation, analysis, prioritization, and network construction as shown in Figure 2.1. Our method expands upon traditional mutation analysis by investigating changes to non-coding regulatory elements. While regulators correspond to a small percentage of total nucleotides, they occur throughout the entire genome and perform functions critical to transcription and translation. EVA prioritizes mutations based on their predicted effect on properties such as transcription and translation completion, structural stability, and binding site affinity. Based on the type of mutation and supporting information from sequence and structure analysis, EVA assigns annotated mutations a priority to aid researchers in interpretation and planning follow-up experimentation. A high priority indicates the mutation is predicted to be damaging to a gene or destroy a regulator function, a low priority indicates the mutation is predicted to be tolerated by the feature, and an unassigned priority denotes an undetermined effect. Finally, EVA identifies features indirectly affected by mutations in genes or regulators (e.g., a mutated promoter indirectly affects genes in its transcription unit) from the *E. coli* gene regulatory network. The *E. coli* metabolic network offers associated reaction and pathway data and insight into the phenotypic impact of mutations.

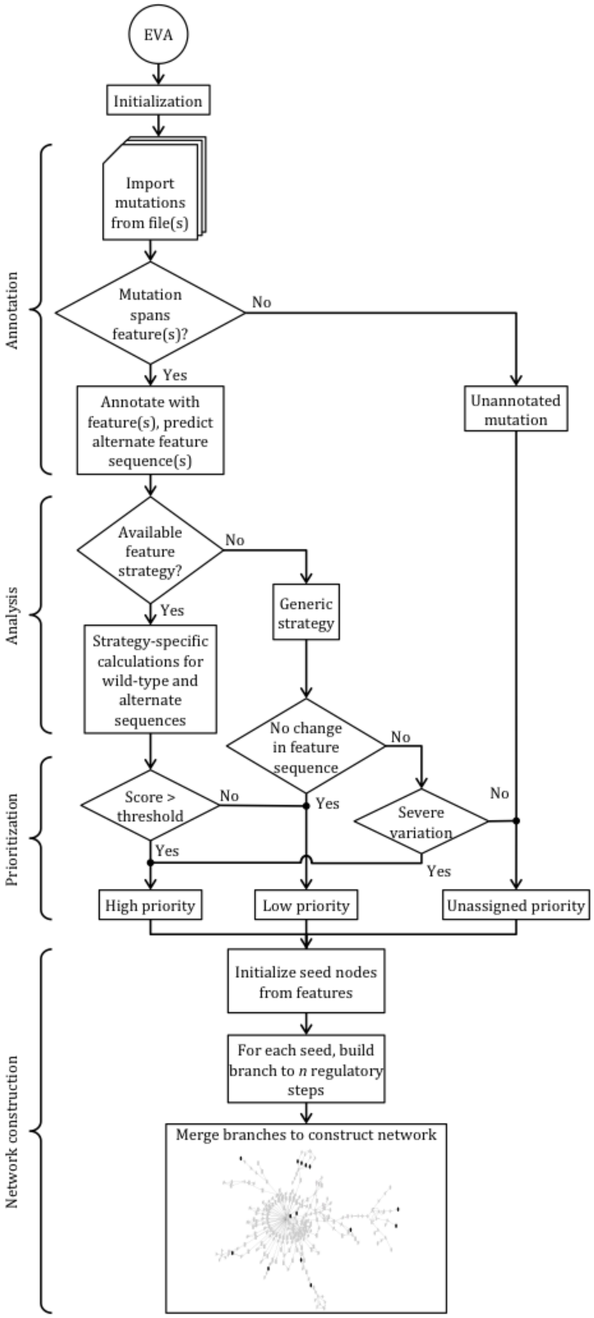

EVA
Initialization
Annotation
Import mutations from file(s)
Mutation spans feature(s)?
No
Yes
Annotate with feature(s), predict alternate feature sequence(s)
Unannotated mutation
Analysis
Available feature strategy?
No
Yes
Generic strategy
Strategy-specific calculations for wild-type and alternate sequences
No change in feature sequence
No
Yes
Prioritization
Score > threshold
No
Yes
Severe variation
No
Yes
High priority
Low priority
Unassigned priority
Network construction
Initialize seed nodes from features
For each seed, build branch to n regulatory steps
Merge branches to construct network

Figure 2.1 *EVA pipeline*

Mutations are imported and are annotated with corresponding genomic features that include coding regions, structures, or binding sites. Each annotation is processed as either DNA or RNA and the reference (wild-type) sequence is compared to the alternate sequence to classify the mutation type. Mutations are classified based on if they are predicted to be damaging to the genomic feature. Nodes corresponding to genomic features in annotations are used as seed nodes to build a biological network based on gene regulatory and metabolic networks that illustrate downstream effects of genomic mutations.

Methods

Mutation Annotation

EVA compiles relevant genomic feature data (e.g., feature type, name, strand, position, and sequence) and determines the span for each annotation. The mutation span describes the mutation and genomic feature intersection on the genome. These cases include when the mutation is internal to the genomic feature, the mutation spans the left or right end of the genomic feature, the mutation coincides exactly with the genomic feature, or the mutation encompasses the genomic feature and surrounding DNA. Pairs of mutations and genomic features together form annotations.

EVA accepts Variant Call Format (VCF version 4.3) files (.vcf), Breseq Genome Diff output files (.gd) or a comma-delimited text file (.csv or .txt) as input. CSV files must have one mutation per line, given as the genomic position, reference (wild-type) DNA sequence, and alternate DNA sequence. Users may submit input files designated as parent (i.e., ancestral) strains or derivative (i.e., evolved) strains. EVA requires at least one derivative strain.

EVA implements an interface to RegulonDB (Gama-Castro et al.) that it uses to execute queries and retrieve annotation data for *E. coli*. EVA supports RegulonDB versions 9.1, 9.2, and 9.3, and 9.4. RegulonDB version 9.1, corresponds to EcoCyc version 19.5 (Keseler et al.) and the *E. coli* K-12 MG1655 genome version U00096.2 (Riley), while later versions correspond to EcoCyc versions 20.0, 20.1, and 20.5, respectively, and *E. coli* K-12 MG1655 genome version U00096.3 (Hayashi et al.). EVA additionally requires a supplementary database derived from EcoCyc 21.0. All accessed databases and short read alignment algorithms should use the same version of the genome for compatibility when referencing absolute positions on the genome. *E. coli* strains that are highly similar to

MG1655 can also benefit from EVA. To do so, researchers must first align short-read data to the reference genome that corresponds to their installation of RegulonDB. EVA output refers to reference genome position coordinates in MG1655, but annotations and analysis will generally be consistent. Inherent differences between MG1655 and an alternate strain will be present across all samples and EVA will not consider these variations during network construction.

For each mutation, EVA queries the RegulonDB database for features that coincide with the specified mutation region. During the annotation step, EVA considers all RegulonDB objects with defined absolute positions on the genome. The RegulonDB database maintains such data for genes, promoters, ribosome binding sites (RBS), terminators, attenuators, sRNA binding sites, riboswitches, and transcription factor binding sites (TFBS) and we refer to these as genomic features. Figure 2.2 gives an overview of several genomic features and their roles in *E. coli*. A mutation may coincide with multiple genomic features or no genomic features. By default, EVA does not annotate regulatory features for which the entire regulon also occurs inside the mutation to avoid network representation of mutated regulators that have no known added effect on the organism when a mutation is large. Users may change this setting to report all features in a region if desired. To reduce computation time for large mutations, EVA will only annotate coding regions. The default threshold for this behavior is 1 kilobase (kb) but users can change this to an alternative size.

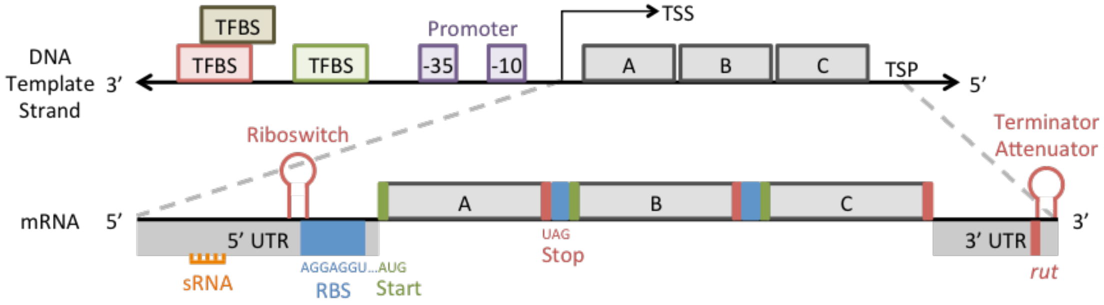


Figure 2.2 *Gene model*

A gene regulation model for a bacterial system. Proteins that bind at TFBS regulate transcription of the DNA template strand. The promoter is responsible for recruiting RNA polymerase and transcription begins at the transcription start site (TSS). Transcription continues through genes A, B, and C until RNA polymerase stalls at the termination stop point (TSP). This may occur due to either a terminating hairpin followed by a U track, or Rho factor, which is recruited at the Rho utilization (rut) site. Translation of mRNA is regulated by antisense sRNAs, riboswitches, and RBSs that contain the Shine-Dalgarno sequence.

Analysis and Prioritization

Automating the analysis and prioritization of mutated features enables investigators to distinguish between mutations predicted to be damaging from mutations predicted to be tolerated. EVA employs different analysis strategies for mutations in coding regions, structural features and RNAs, and binding sites, but the design allows for additional methods to be incorporated in the future. This section presents the implemented strategies for analyzing and prioritizing annotations for supported genomic features.

Genes

EVA classifies mutations that correspond to genes that encode proteins by the resulting change in the amino acid sequence and assigns high or low priorities based on the predicted severity of impact as outlined in **Error! Reference source not found.**. The reference nucleotide sequence refers to the wild-type gene sequence in the reference genome. From this sequence, EVA derives an alternate nucleotide sequence by substituting the corresponding region in the reference sequence with the alternate mutation sequence. In some cases, constructing the alternate sequence requires additional processing. For example, in the case of a frameshift or deletion of the stop codon, the alternate gene sequence is the DNA sequence that begins at the gene start codon and extends to the first recognized stop codon. EVA classifies mutations that encompass a gene or modify the start codon without resulting in an alternate start codon as knockouts (KO) or loss of start codon and does not perform further analysis.

Table 2.1 *Prioritization of mutations in protein-coding genes*

Provean scores ≤ -2.5 are predicted to be damaging and are assigned a high priority. Scores not meeting the cutoff are predicted to be tolerated. In cases where a score cannot be computed, the priority is undefined and the mutations require further review by the investigator.

Variation	Description
Low priority	
Silent	No change in amino acid sequence.
Alternate Stop	Stop codon is replaced with an alternate stop codon.
High priority	
KO	Entire gene has been deleted or altered.
Alternate Start	Start codon is replaced with an alternate start codon.
Start Loss	Loss of start codon.
Nonsense	Substitution of one amino acid for a stop codon.
Determined by Provean score	
Missense	Substitution of one amino acid.
Deletion	Deletion of one or more amino acids.
Insertion	Insertion of one or more amino acids.
Delins	Deletion followed by insertion of one or more amino acids.
Frameshift+	Out-of-frame insertion.
Frameshift-	Out-of-frame-deletion.
Read-through	Loss of stop codon results in read-through to next stop codon.
Duplication	Duplication of an amino acid sequence.

Prioritization of mutations that result in variant amino acid sequences is based on Provean scores. We compiled libraries of supporting sequences for the *E. coli* U00096.2 and U00096.3 transcriptomes to improve performance and avoid the time-consuming homology search for each gene annotation. Libraries were generated using Provean version 1.1.5, NCBI BLAST version 2.4.0+, CD-HIT version 4.6.4, and the NCBI BLAST non-redundant sequence database (last updated on January 12, 2015) using Cyverse resources (Merchant et al.).

EVA translates both the reference and alternate nucleotide sequences to amino acid sequences using the bacterial genetic code specified in NCBI translation table 11. If there exists no difference between the reference and alternate amino acid sequences, the gene mutation is silent. Where possible, EVA produces variation descriptions using HGVS nomenclature from the reference and alternate amino acid sequences. EVA generates this description by calculating the greatest common prefix and greatest common suffix of the reference and alternate amino acid sequences, assessing the sequence variation, and selecting the proper HGVS descriptor for the change in amino acid sequences.

Because Provean only supports certain HGVS formats as input, EVA classifies amino acid variations as missense (single amino acid substitution), nonsense (premature stop codon), insertion (insertion of one or more amino acids), deletion (deletion of one or more amino acids), delins (deletion followed by an insertion), or duplication (duplication of amino acid region). EVA converts other mutations, such as frameshifts, into delins when possible. The reference amino acid sequence and HGVS variation description are submitted for Provean analysis.

EVA prioritizes gene annotations based on mutation type and Provean results. Silent mutations and those predicted to be tolerated (default threshold is a Provean score > -2.5) receive low priority, as they are unlikely to affect protein function. KOs, loss of start codons, and mutations predicted to be damaging are most likely to destroy protein function and are assigned high priority. Nonsense mutations and frameshifts are typically given a high priority unless they occur toward the end of the coding sequence. In some cases, such as a low number of sequence homologs, the Provean analysis may not be relevant, and EVA regards the priority as unassigned.

Structural Features and RNAs

Annotations corresponding to terminators, attenuators, riboswitches, and RNA genomic features are analyzed for changes in secondary structure. The reference nucleotide sequence refers to the wild-type genomic feature sequence in the reference genome. EVA constructs an alternate nucleotide sequence by substituting the region in the reference sequence that corresponds to the mutation with the alternate mutation sequence. The RNAfold methods in the Vienna RNA package (Hofacker, 2003) predict the secondary structure and calculate minimum free energy (MFE) for the reference and alternate sequences. The RNAfold mfold algorithm uses dynamic programming to predict an energetically stable model of an RNA molecule by minimizing its free energy. The energy model sums contributing free energies from loops to calculate the total free energy score of a secondary structure.

Comparing the MFE (in kcal/mol) for reference and alternate sequences reveals if the mutation affects the secondary structure stability; a smaller MFE in the alternate sequence indicates greater secondary structure stability while a larger MFE indicates reduced secondary structure stability. For example, a mutation in a terminator that reduces MFE of the predicted secondary structure suggests a higher likelihood of forming its stem-loop structure and stopping transcription. To offer context for the change in MFE, we simulate variations of the reference sequence and calculate the MFE for each variation. Variant sequences include single nucleotide deletions, insertions, and substitutions at each position. With insight into the energy landscape of the molecule subjected to small variations, we can compare the change in MFE caused by the mutation with other minor sequence changes. Calculating the Levenshtein distance (Levenshtein), or another metric such as a mutual information score, for aligned structures in dot-bracket notation captures the change in predicted secondary structure.

A change in predicted secondary structure may impair or destroy the function of regulatory elements or disrupt protein folding. Thus, if the change in MFE exceeds a user-defined cutoff (e.g., greater than 1 standard deviation from reference) or the Levenshtein distance exceeds a threshold, EVA assigns annotations a high priority. EVA assigns a low priority to annotations corresponding to mutations that result in small changes in MFE or do not significantly alter the predicted secondary structure. Because computational requirements for secondary structure prediction grow exponentially with sequence length, EVA limits predictions to sequences with length less than 1 kb, but users may override this setting or independently run predictions. If secondary structures are not predicted, EVA considers the priority to be unassigned.

Binding Sites

EVA searches relevant genomic feature sequences to ascertain if the alternate sequence of a mutated feature is a known binding site sequence in *E. coli*. This is performed by querying the RegulonDB database for unique sequences with the same function (e.g., all σ^{28} promoters or all TFBS to which the transcription factor Fis is predicted to bind). If the alternate sequence is a known binding site, while the mutation may change binding site affinity, EVA predicts it to remain functional and assigns a low priority. Otherwise, EVA assigns the annotation a high priority.

For σ^{70} promoters, the most prevalent class of promoters and the class associated with the primary sigma factor during exponential growth, EVA may perform an alternative analysis. Kinney et al. developed a procedure known as Sort-Seq to create a predictive map for the *E. coli* RNAP as they bind to DNA (Kinney, Murugan, Callan, & Cox). Their experiment characterized over 200,000 variations on a 75 bp region of the lac promoter and CRP binding site and authors inferred energy matrices that described the CRP-DNA and RNAP-DNA interactions.

Despite the complexity of protein-DNA interactions, it has been shown that a sequence-dependent linear model sufficiently describes binding energy for DNA-protein interactions (Benos). Each base in the DNA sequence is modeled as having an independent contribution to overall binding affinity. Thus, given parameters defined in (Kinney et al.), the binding energy of RNAP to a specific DNA sequence is the sum of energy values from contributing bases along the sequence.

EVA annotates the -10 and -35 elements of a promoter as separate genomic features, thus Sort-Seq scores are calculated independently for each promoter element. EVA only scores substitutions in promoter elements that are the same size as those in the Sort-Seq matrix, 6 bp, which is also the predominant promoter element size. A specific promoter element size is not a limitation of the Sort-Seq experimental method, and if alternate sequence lengths were measured, EVA could appropriately penalize insertions and deletions inside promoter elements. For a substitution in a σ^{70} promoter feature that is 6 bp, EVA assesses if the Sort-Seq score for the reference (wild-type) promoter differs from the alternate promoter sequence resulting from a mutation. Following the assumption that gene expression is proportional to the probability that RNAP is bound, it follows that an increase in binding affinity will result in an increase in mRNA abundance and a decrease in binding affinity will result in a decrease in mRNA abundance. If there is no change, EVA classifies the annotation as low priority, otherwise, EVA assigns a high priority.

Gene regulatory and metabolic network generation

Annotations and mutation analysis only suggest direct effects of changes in the genome. EVA utilizes gene regulatory and metabolic network data to visualize downstream features to assess the indirect effects a mutated genomic feature can have on regulated features. EVA generates a biological network by retrieving downstream features of the mutated feature via transcription and translation regulation and signaling and metabolic pathways. The expanded collection of elements forms a gene regulatory and metabolic network. In this network, nodes are biological features such as genomic features, gene products, transcription factors, transcription units, reactions, and pathways (Figure 2.3). Directional edges denote interactions such as transcription regulation, translation regulation, and catalysis or relationships such as a gene encoding a protein, a transcription unit comprising genes, or a reaction belonging to a pathway. Clustered nodes can show relationships between mutations and biological systems relevant to organism fitness.

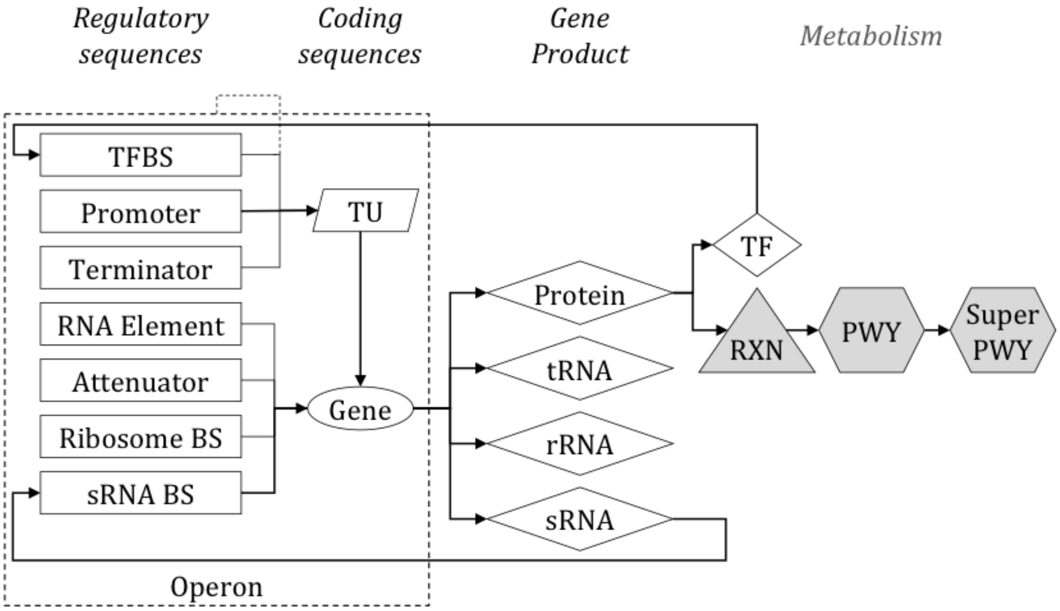


Figure 2.3 *Combined gene regulatory and metabolic network structure*

Nodes representing regulators that map to absolute positions on the genome are represented in rectangles (e.g., promoters, terminators). Genes are represented by ovals, transcription units (TU) by parallelograms, gene products by diamonds, reactions (RXN) by triangles, and pathways (PWY) by hexagrams. Mutations are annotated with operons only if no other annotation is available. The operon region spans the transcription units as well as regulators of the transcription units. Directed edges link nodes based on regulatory activities, transcription, translation, and enzymatic activity. Gene regulatory data are obtained from the RegulonDB database and reaction and pathway data, shaded in the figure, are obtained from EcoCyc. The default EVA network bypasses TU features and connects regulators directly to genes. TF and sRNA nodes bypass binding sites in favor of edges to regulated genes. TFBS and sRNA BS are only represented if they have an associated mutation.

For network construction, the default behavior is to convert annotations into a set of genomic features, each with an associated list of strains. When a user submits more than one strain for analysis, EVA ignores features that contain mutations in all strains during network construction. This is done to eliminate background variations from the network for strains highly similar to *E. coli* K-12 MG1655 and variations propagated from ancestral strains.

EVA additionally gives an option to produce networks for each derivative sample with variations common to ancestral strains removed.

For each genomic feature, EVA creates a seed node representing the feature and initializes a branch, a directed, rooted network which may contain cycles, with the seed node as the root. To build the branch, EVA adds outward edges and nodes recursively until a user-defined number of transcription and/or translation regulation steps, *n*, is reached (default *n* = 1). Network construction does not penalize other types edges (i.e., relationship designations between nodes, such as a reaction belonging to a pathway) in this process. If a node is a regulator, biological features controlled by the regulator and interaction edges are added to the network. If the node is a gene, the gene product, associated enzymatic reactions and associated biological pathways are added to the network along with appropriate edges. If the node is a transcription unit, genes within the transcription unit are added to the network with edges connecting the transcription unit to the genes. This process continues up to *n* regulation steps.

EVA constructs a branch for each genomic feature with one or more associated mutations. Finally, EVA merges all branches into a single network that represents mutated elements and their downstream features. EVA, by default, will generate a network that merges some linear relationships to reduce network size, however an option to represent all features is available. In addition, EVA produces two alternative networks to aid in biological interpretation: a mutation interaction network, and a shortest-paths network. The default network produced by EVA contains all biological features that can be reached within *n* regulation steps of a variant genomic feature, where *n* is a parameter specified by the user at runtime. This network represents all potential downstream biological features that one or more mutations may affect and which may contribute to the phenotype of a derivative strain.

The mutation interaction network is a subset of the default network formed of all paths from seeds to nodes reachable by two or more seeds. Thus, leaf nodes and most internal nodes are biological features that multiple mutations may affect indirectly. While mutations may individually yield a specific phenotype, this network can reveal potential interactions among mutations which could be synergistic, additive, or antagonistic. The shortest-paths network is a minimal representation of potential interactions among seeds. EVA constructs this network from all shortest paths, measured by the number of edges, from pairs of seed nodes to common nodes among branches. This visualization offers a minimal summary of EVA results which diminishes contributions of large regulators, such as transcription factors, which can overwhelm the network. Every network depicts all seed nodes even if they are isolated with degree zero. A single node attribute file may be used for all representations. EVA writes attributes and network files to files that can be imported into Cytoscape (Shannon, 2003) for visualization. A Cytoscape style has also made available with EVA.

Software implementation

The EVA core software was developed in Java. A local copy of RegulonDB with supplemental tables utilizing metabolic network information from EcoCyc was used. A PROVEAN library for *E. coli* genes was created for faster mutation analysis. Source code is available under an open source license at https://github.com/eboggess/EVA.

Results

For the following experiments, short read data alignment to *E. coli* K-12 MG1655 (U00096.3) and SNP-calling was performed using Breseq 0.31.0 (Deatherage & Barrick) with Bowtie2 2.3.2 (Langmead & Salzberg) and R 3.4.1. A local installation of RegulonDB 9.4 was used to annotate mutations and retrieve biological feature interactions for network construction. A local copy of a supplemental database derived from EcoCyc 21.0 was used to incorporate reaction and pathway data into biological networks. Secondary structure analysis was performed with the Vienna RNA package 2.3.2 and amino acid sequence variations were score using Provean 1.1.5 with CD-HIT 4.6.4 and NCBI blast 2.4.0+. All ancestral and derivative strain mutation data was submitted to EVA as Genome Diff files and default options were used in the analyses.

Antibiotic resistance in *E. coli*

This section compares the genomic mutations identified and analyzed in work by Wang et al. (Wang et al.) with the expanded analysis provided by EVA. In the metabolic evolution experiment performed by Wang et al., *E. coli* K-12 MG1655 was used as the parent strain and exposed to antibiotics with the goal of generating a strain that exhibits antibiotic tolerance and identifying mechanisms of drug resistance. Fifty parallel populations of the ancestral strain were exposed to antibiotics Ciprofloxacin (Cpr), Neomycin (NeoB), a Cpr-Neo hybrid (Hyb), a Cpr/NeoB equipolar mixture (EqP), and a Cpr/NeoB equimolar mixture (EqM). Four parallel populations were grown with no evolutionary pressure as a control (Ctrl). At the conclusion of the experiment, genome sequencing was performed, short read data were aligned to the *E. coli* K-12 MG1655 (U00096.2) genome (Riley) with Bowtie2, and mutations were predicted using SAMtools and Dindel (Albers et al.). One hundred eighty-four unique mutations corresponding to 93 genes and 5 tRNAs were predicted in one or more strains evolved for drug resistance and 6 mutations were annotated with the nearest gene.

We applied the EVA pipeline to the genomic data from Wang et al., beginning with short-read alignment and SNP-calling with Breseq. Raw SNP-calling results include instances where only reads aligned in one direction support an alternate base call. This may be an artifact from errors in library preparation or sequencing, or an error in the alignment process (Guo et al.). Breseq uses a Fisher's exact test for biased strand distribution and a Kolmogorov-Smirnov test for lower quality reads supporting the alternate sequence to reduce false positive SNP calls. Our analysis using Breseq can detect more complex variations, such as large deletions, and is more accurate in more accurate in finding mutations. In total, Breseq predicts 232 unique mutations among the 54 samples, including 182 of the mutations predicted by Wang et al. The 50 new mutations correspond primarily to large insertions and deletions the earlier method may not detect, but 16 are previously unreported single nucleotide variants (SNV) and small insertions or deletions (less than 10 bp). Breseq detects the reported 2 bp insertion in *yqgE* but excludes it from analysis due to low-quality base calls for the alternate sequence and strand bias in the alignment. Discrepancies exist between the results, including a predicted 7 bp insertion after *dnaG* not in our results. Wang et al. predict a 3 bp deletion (TGG) at relative position 1787 of 2145 in *pta*, but our results instead have a 3 bp deletion in *pta* at relative position 1789. Finally, our results do not have the SNV in *fusA* that Wang et al. predicts for six samples in the NeoB-08 sample.

Breseq Genome Diff output files for the ancestral and derivative strains were provided to EVA as input for mutation analysis. EVA produces 266 annotations for the predicted mutations, meaning that some mutations spanned more than one genomic feature. EVA reports 17 annotations that correspond to regulators not represented in Breseq output and annotations corresponding to insertion sequences, pseudo genes, and operons for five mutations. Table 2.2 provides a summary of selected results and full details are available in an additional file.

Table 2.2 *A selection of previously unreported predicted mutations in antibiotic-resistant strains*

Annotations are listed alphabetically by name along with the corresponding antibiotic treatment(s) and assigned priority are provided for each annotation. The acrA attenuator, gntR terminator, sulA promoter, and LexA binding site are EVA annotations not provided by Breseq output.

Annotation	Antibiotic(s)	Priority
***acrA* attenuator**	**HyB**	**High**
cra	Hyb	High
cyoE-ampG	NeoB	High
emrR	EqM	High
fre	NeoB	High
ftsZ	EqM	Low
***gntR* terminator**	**EqM**	**Low**
icd	Cpr, EqM, Hyb	Low
lexA	HyB	High
nikA	EqM	High
nuoC	NeoB	High
rhsC	EqP	Low
rimK, *ybjN*, *potF*, *potG*, *potH*, *potI*	Cpr	High
rrlC	EqP	Undefined
sucD	HyB	High
***sulAp* / LexA TFBS**	**EqP**	**Undefined**
tamA-tamB	EqP	High
tufB	EqP	High
waaQ	Hyb, EqP	High
yaiU-[*yaiW*]	NeoB	High
yqjI	EqM	High

EVA excluded genomic features with any mutation in all 54 samples from network construction. The gene regulatory and metabolic network for the experiment is built from seed nodes that represent the remaining 164 features. Network summary statistics are provided in **Error! Reference source not found.**. Network size was reduced by more than half in the mutation interaction network, primarily due to the exclusion of the LexA regulon. Cytoscape was used to visualize the simplified shortest-paths network shown in Figure 2.4A.

Table 2.3 *Summary of nodes and edges in biological networks generated by EVA.*

The default network includes all nodes and edges. The mutation interaction is a subset of the default network comprised of nodes representing mutated features and edges that connect them. The shortest path network further reduces the number of nodes and edges by including only the shortest paths between mutated features.

Network	Nodes	Edges
Antibiotic-resistant (161 seed nodes)		
Default	1122	1329
Mutation interaction	503	621
Shortest Path	325	270
Benzoate tolerant (104 seed nodes)		
Default	797	1008
Mutation Interaction	437	598
Shortest path	234	200

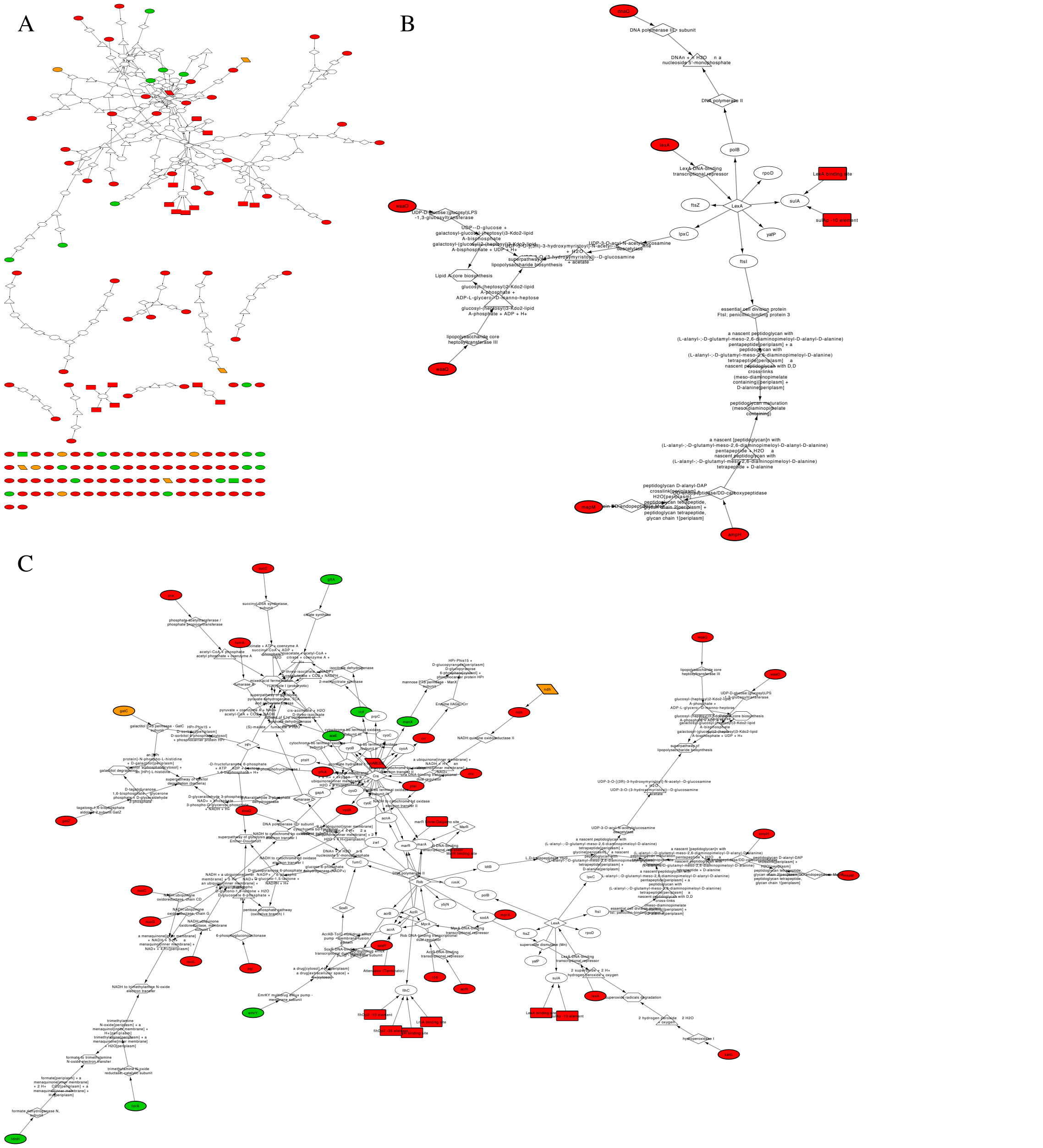


Figure 2.4 *Network visualization of mutations in antibiotic resistant strains*

A. Cytoscape visualization of the EVA shortest-paths network (325 nodes, 270 edges) derived from predicted variant genomic features in antibiotic-resistant strains. Red indicates seed nodes representing features assigned a high priority, orange indicates unassigned priority, and green indicates low priority. The network includes up to 1 level of transcription or translation regulation downstream of a mutated feature. B. The largest cluster in the antibiotic resistance network is the Cra regulon, marRAB operon, and the multidrug effux system. C. A selection from the cluster in B that feature the LexA regulon and SOS response system. This subset is connected to the remainder of the cluster in B via the peptidoglycan maturation pathway node.

Cra regulon, *marRAB* operon, and multidrug efflux system

Breseq detected evidence of an unreported 12 bp deletion within the *cra* gene at position 88,827 in the Hyb-10 sample. The predicted mutation results in an out of frame deletion beginning at amino acid position 267 of 335 in Cra and is represented in HGVS-style notation as D267_I270del. Due to the severity of the deletion (Provean score -15.83), EVA predicts the mutation to be damaging to Cra, a transcriptional dual regulator. Examining the gene regulatory network shows Cra is a predicted repressor of the *marRAB* operon. Previously, Wang, et al. hypothesized that the hybrid drug could evade the *marRAB* drug efflux, which would make it unique among quinolone drugs. While the predicted *cra* mutation occurs in only one hybrid strain, PCR verification of the deletion and measuring transcript abundance of *marRAB* genes could reveal if regulation affects the operon and if this is a strategy for hybrid drug resistance.

Examination of all predicted mutations in the gene regulatory and metabolic network illustrates how both Cra and AcrR, a *marRAB* repressor corresponding to mutations reported by Wang et al. regulate the *marRAB* operon. Instances of mutations in *acrR* appear in all strains except those evolved for NeoB resistance. EVA provides a second annotation for a SNV within *acrR* that corresponds to an *acrA* attenuator (terminator) (C to A at position 485,010 in strain Hyb-09). The MFEs of the reference and alternate sequences sequence -5.70 and -0.30 kcal/mol, respectively, showing a lower likelihood for the terminator structure to form which would result in an increase in *acrA* transcription. Provean predicts the mutation in AcrR to be damaging with a score of -4.60. Damaged AcrR could result in weaker repression of the *marRAB* operon and increased abundance of the *marRAB* transcript. Because this mutation appears in many strains, including one Hyb strain, there is strong support that altering the AcrR transcription factor is a strategy for antibiotic resistance.

Network visualization in Cytoscape shows the relationship between these regulators and two other multidrug resistance genes; *mprA* and *emrY* (Figure 2.4B). The *emrY* mutation, which is silent, was previously identified in EqP-02, but the 69 bp deletion that affects *mprA* (also known as *emrR*) had not been reported. MprA is a transcriptional repressor that is predicted to regulate the *acrAB* operon. Damage to MprA could decrease repression of *acrAB* and subsequently increase drug transmembrane transport.

LexA regulon

LexA is a transcriptional repressor responsible for regulating the SOS response, the cellular response to DNA damage or inhibition of DNA replication (Janion). EVA assigns the point mutation Wang et al. reported within *lexA* a high priority based on the Provean score of -7.63 which predicts the corresponding P107Q amino acid variation to be damaging. The SOS response can promote mutations, which increases the opportunity to acquire antibiotic resistance (Cirz et al.). By damaging the LexA repressor, the SOS pathway may be de-repressed, enabling increased transcription of error-prone SOS-regulated polymerases.

In addition to the LexA mutation, which EVA predicts to be damaging with a Provean score of -7.61, a previously unreported SNV in EqP-09 (A to G at position 1,020,956) corresponds to both the -10 element of the *sulA* promoter and a LexA binding site (Figure 2.4C), which acts as a transcriptional repressor for *sulA*. SulA is a cell division inhibitor which has been shown to be involved in stress-induced point mutations (McKenzie, Harris, Lee, & Rosenberg).

An unreported 12 bp deletion inside *ftsZ* in EqM-06 results in an amino acid sequence change described by EVA as P346_Q349del. FtsZ, which is essential for cell division, is a known antibiotic target and is inhibited by SulA (Cordell, Robinson, & Lowe). Despite deleting four amino acids from FtsZ, Provean predicts the mutation to be tolerated with a score of -0.599 and EVA assigns a low priority. While predicted mutations in LexA and its regulon occur in distinct strains adapted for under different conditions, the SOS response system may be a relevant antibiotic resistance strategy for both drug mixtures and the hybrid drug. Further experimentation is necessary to examine the variant *sulA* promoter strength and changes in *sulA* and *ftsZ* transcripts and discern relevance to fitness.

Benzoate tolerance in *E. coli*

In this section, we compare genomic mutations identified and for benzoate-adapted strains (Creamer et al.) with the analysis provided by EVA. The benzoate evolution experiment performed by Creamer et al., used *E. coli* W3110 as the parental strain and 24 cultures were subjected to increasing benzoate concentrations up to 20 mM. The Illumina MiSeq platform was used for genomic sequencing of 16 benzoate-evolved strains and the parent strain. Creamer et al. used Breseq to assemble short-reads and identify of genomic variants. 110 mutations were predicted in one or more evolved strain, but not the parent strain.

For our analysis, we repeated the short-read assembly and annotation using Breseq and used *E. coli* K-12 MG1655 U00096.3 as the reference genome for compatibility with EVA. Because the strains are highly similar, the 136 mutations predicted by Breseq are generally consistent with those reported by Creamer et al., but with MG1655 genome position coordinates. From these 136 mutations, EVA finds 188 corresponding genomic features, of which 104 features are not mutated in all strains.

In addition to results reported by Creamer et al., EVA annotates nine mutations previously only recognized as intergenic and adds regulatory feature annotations to three reported gene mutations (Table 2.4). The network generated by EVA links the missense mutation in rob with the *marRAB* operon which is deleted from the genome in several strains. Genes *cpxA*, *emrY*, and *emrA* appear separately as they are part of a signal transduction system and efflux pumps. The Provean score for the L191M amino acid mutation in MdtA does not meet the minimum criteria and EVA considers it a low priority mutation. For one mutation in gene *fdnG*, a Provean analysis was not performed due to a selenocysteine site which appears as an internal stop codon during translation; a scenario for which EVA is unable to provide analysis.

Creamer et al. noted there must be factors besides those reported which are responsible for fitness advantage and chloramphenicol sensitivity based on their strain characterizations. Specifically, the authors hypothesize that the C3-1 genome may have defects in other multidrug resistance genes, G5-2 must have unknown mutations that contribute to chloramphenicol sensitivity and benzoate fitness, and E1-1 maintains chloramphenicol resistance, unlike other strains. To further investigate variations in the phenotype among evolved strains, we examine the predicted mutations in the combined gene regulatory and metabolic network representation generated by EVA.

Table 2.4 *Previously reported intergenic mutations in one or more benzoate-tolerant strains for which EVA provides annotations. Mutations are ordered by genome position.*

Mutation	Previously reported annotation	EVA annotation	Priority
29617, A → G	intergenic (+422/-34), *dapB* → / → *carA*	ArgR TFBS (repressor)	Undefined
573671, T → A	intergenic (+109/+289), *ybcQ* → / ← *insH*	*ipeX*	Undefined
1337160, G → A	intergenic (+617/-385) *cysB* → / → *acnA*	*yciX*	Low
1553904, 2 bp → CT	intergenic (+199/+207), *fdnI* → / ← *yddM*	*C0362*	Low
1553926, T → C	intergenic (+199/+207), *fdnI* → / ← *yddM*	*C0362*	Low
1565001, A → G	intergenic (-211/+47), *ddpX* ← / ← *dos*	Rho-independent terminator	Low
1908956, IS5 (–) +4 bp	coding (191-194/210 nt), *cspC* ←	rlmA Attenuator (anti-terminator)	High
1908956, IS5 (–) +4 bp	coding (191-194/210 nt), *cspC* ←	Riboswitch	High
1909258, IS1 (+) +9 bp	coding (40-48/144 nt), *yobF* ←	Riboswitch	High
2441649, C → T	intergenic (-44/-115), *fabB* ← / → *trmC*	*fabBp*	Undefined
4218986, IS5 (–) +4 bp	intergenic (+187/-79), *metA* → / → *aceB*	*aceBp*	High
4470927, G → A	intergenic (-67/+52), *treB* ← / ← *treR*	*treB* Attenuator (terminator)	Undefined
4639891, A → G	S34P (TCC→CCC) *rob* ←	(in addition to rob) *creAp*	Low

Several regulators are predicted to contain mutations, including a *dosCP* terminator, *fabB* promoter, and *aceBAK* promoter. The point mutation in the *dosCP* terminator in the E1-1 clade is interesting because it is upstream of *ddpX*, a D-Ala-D-Ala dipeptidase involved in resistance to antibiotic vancomycin (Lessard & Walsh). However, the resulting sequence change does not affect MFE of the predicted secondary structure and transcriptomic analysis of *ddpX* is required to determine if the mutation has any effect. The point mutation in the *fabB* -10 promoter element in all strains in the G5-2 clade results in a more favorable sequence for σ^{70} binding. An insertion sequence is detected inside the -10 promoter element for *aceBAK* in the E1-1 clades. Separately, a point mutation that is predicted to be damaging is found in *aceA* in all strains in the C3-1 clades (Figure 2.5). These mutations may represent different strategies to manipulate glyoxylate metabolism in benzoate adapted strains.

EVA annotates a mutation previously reported in an intergenic region with the small RNA *ipeX*. The point mutation in *ipeX* in the A5-1 clades results in a higher MFE and thus a less favorable secondary structure. The small RNA *ipeX* has been shown to inhibit expression of outer membrane porins *ompC* and *ompF* through post-transcriptional modification (Castillo-Keller, Vuong, & Misra).

EVA identifies a relationship between previously reported mutations in the genes *add* and *deoD*. The gene *add* contains a frameshift mutation in all strains in the G5-2 clade and *deoD* contains a predicted damaging SNP in G5-1. These genes are connected by the purine salvage pathway (Figure 2.5). Separately, strains in the E1-1 clade contain a mutation predicted to be damaging in the *apt* gene, which encodes another purine salvage enzyme. Purine metabolism is affected by antibiotics and has been proposed as a drug target for resistant bacteria (Møller et al.). Network statistics are provided in Table 2.3.

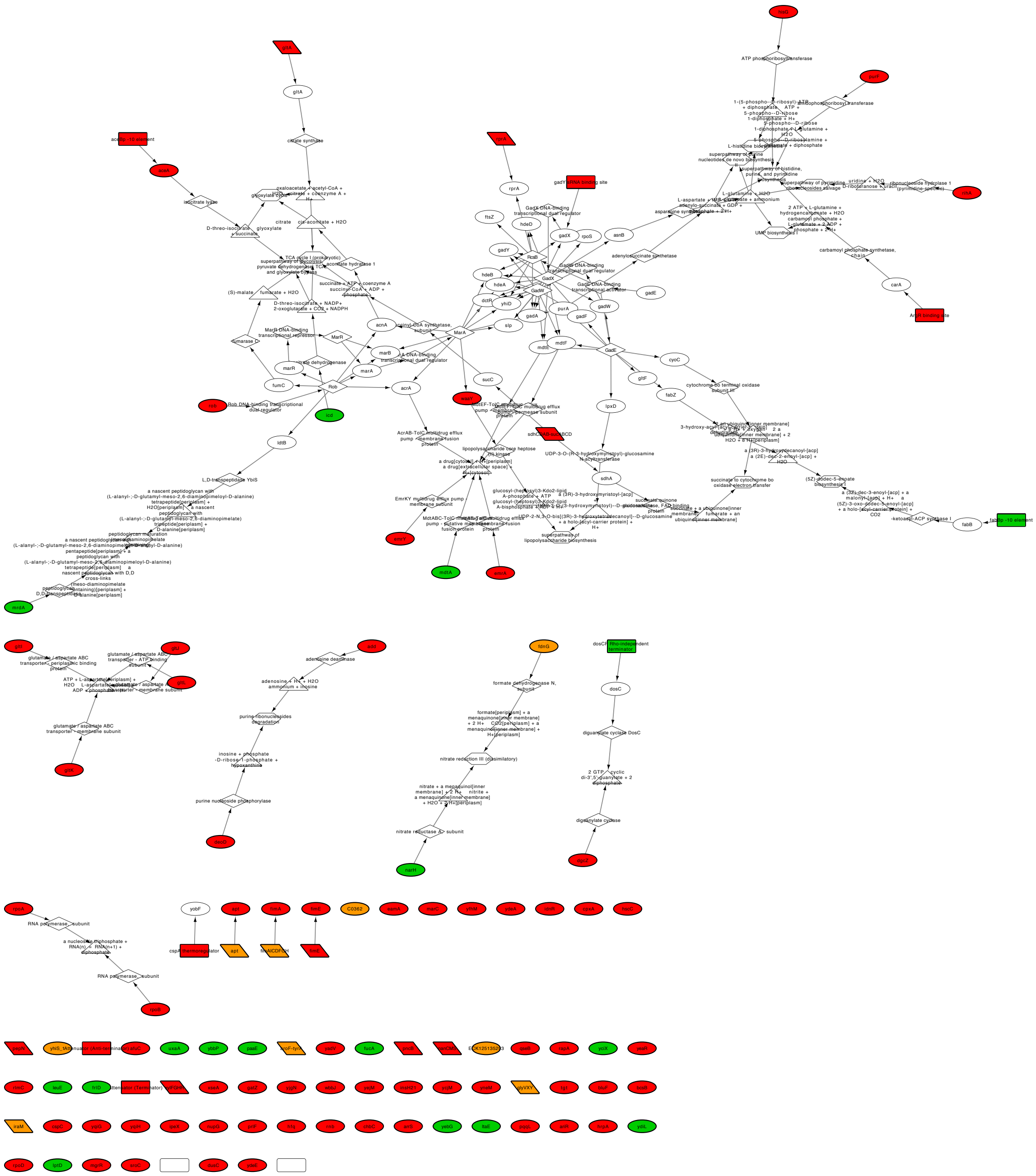

gltA
citrate synthase
aceBp -10 element
aceA
isocitrate lyase
glyoxylate cycle
oxaloacetate + acetyl-CoA + H2O
citrate + coenzyme A + H+
D-threo-isocitrate
succinate
glyoxylate
citrate
cis-aconitate + H2O
TCA cycle I (prokaryotic)
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
aconitate hydratase 1
succinate + ATP + coenzyme A
succinyl-CoA + ADP + phosphate
(S)-malate
fumarate + H2O
D-threo-isocitrate + NADP+
2-oxoglutarate + CO2 + NADPH
fumarase C
MarR DNA-binding transcriptional repressor
MarR
marR
isocitrate dehydrogenase
marB
marA
MarA DNA-binding transcriptional dual regulator
acnA
succinyl-CoA synthetase, subunit
MarA
fumC
Rob
rob
Rob DNA-binding transcriptional dual regulator
icd
acrA
sucC
waaY
ldtB
L,D-transpeptidase YbiS
a nascent peptidoglycan with (L-alanyl-;-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanine) tetrapeptide[periplasm] + H2O[periplasm]
a nascent peptidoglycan with (L-alanyl-;-D-glutamyl-meso-2,6-diaminopimelate) tripeptide[periplasm] + D-alanine[periplasm]
peptidoglycan maturation (meso-diaminopimelate containing)
a nascent peptidoglycan with (L-alanyl-;-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine) pentapeptide[periplasm] + a peptidoglycan with (L-alanyl-;-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanine) tetrapeptide[periplasm]
a nascent peptidoglycan with D,D cross-links (meso-diaminopimelate containing)[periplasm] + D-alanine[periplasm]
peptidoglycan D,D-transpeptidase
mrdA
rprA
gadY sRNA binding site
GadX DNA-binding transcriptional dual regulator
ftsZ
hdeD
gadX
rpoS
asnB
gadY
RcsB
hdeB
hdeA
GadX
GadW
GadW DNA-binding transcriptional dual regulator
GadE DNA-binding transcriptional activator
gadE
dctR
yhiD
gadA
purA
gadW
gadF
slp
mdtE
mdtF
GadE
cyoC
gltF
fabZ
lpxD
MdtEF-TolC multidrug efflux pump - membrane fusion protein
MdtEF-TolC multidrug efflux pump - permease subunit
sdhCDAB-sucABCD
AcrAB-TolC multidrug efflux pump - membrane fusion protein
lipopolysaccharide core heptose (II) kinase
a drug[cytosol] + H+[periplasm]
a drug[extracellular space] + H+[cytosol]
EmrKY multidrug efflux pump - membrane subunit
MdtABC-TolC multidrug efflux pump - putative membrane fusion protein
emrY
mdtA
emrA
glucosyl-(heptosyl)3-Kdo2-lipid A-phosphate + ATP
glucosyl-(heptosyl)3-Kdo2-lipid A-bisphosphate + ADP + H+
UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine N-acyltransferase
sdhA
a (3R)-3-hydroxymyristoyl-[acp]
UDP-3-O-(3-hydroxymyristoyl)--D-glucosamine
UDP-2-N,3-O-bis[(3R)-3-hydroxytetradecanoyl]--D-glucosamine + a holo-[acyl-carrier protein] + H+
superpathway of lipopolysaccharide biosynthesis
succinate:quinone oxidoreductase, FAD binding protein
succinate + a ubiquinone[inner membrane]
fumarate + an ubiquinol[inner membrane]
succinate to cytochrome bo oxidase electron transfer
cytochrome bo terminal oxidase subunit III
3-hydroxy-acyl-[acyl-carrier-protein] dehydratase
2 an ubiquinol[inner membrane] + oxygen
2 a ubiquinone[inner membrane] + 2 H2O + 8 H+[periplasm]
a (3R)-3-hydroxydecanoyl-[acp]
a (2E)-dec-2-enoyl-[acp] + H2O
(5Z)-dodec-5-enoate biosynthesis I
a (3Z)-dec-3-enoyl-[acp] + a malonyl-[acp] + H+
(5Z)-3-oxo-dodec-5-enoyl-[acp] + a holo-[acyl-carrier protein] + CO2
-ketoacyl-ACP synthase I
fabB
fabBp -10 element
hisG
ATP phosphoribosyltransferase
purF
amidophosphoribosyl transferase
1-(5-phospho--D-ribosyl)-ATP + diphosphate
ATP + 5-phospho--D-ribose 1-diphosphate + H+
5-phospho--D-ribose 1-diphosphate + L-glutamine + H2O
5-phospho--D-ribosylamine + L-glutamate + diphosphate
L-histidine biosynthesis
superpathway of purine nucleotides de novo biosynthesis II
superpathway of histidine, purine, and pyrimidine biosynthesis
superpathway of pyrimidine ribonucleosides salvage
uridine + H2O
D-ribofuranose + uracil
ribonucleoside hydrolase 1 (pyrimidine-specific)
rihA
L-glutamine + H2O
L-glutamate + ammonium
L-aspartate + IMP + GTP
adenylo-succinate + GDP + phosphate + 2 H+
asparagine synthetase
adenylosuccinate synthetase
UMP biosynthesis I
2 ATP + L-glutamine + hydrogencarbonate + H2O
carbamoyl phosphate + L-glutamate + 2 ADP + phosphate + 2 H+
carbamoyl phosphate synthetase, chain
carA
ArgR binding site
gltI
glutamate / aspartate ABC transporter - periplasmic binding protein
gltJ
glutamate / aspartate ABC transporter - ATP binding subunit
ATP + L-aspartate[periplasm] + H2O
L-aspartate[cytosol] + ADP + phosphate + H+
gltL
glutamate / aspartate ABC transporter - membrane subunit
gltK
add
adenosine deaminase
adenosine + H+ + H2O
ammonium + inosine
purine ribonucleosides degradation
inosine + phosphate
-D-ribose-1-phosphate + hypoxanthine
purine nucleoside phosphorylase
deoD
fdnG
formate dehydrogenase N, subunit
formate[periplasm] + a menaquinone[inner membrane] + 2 H+
CO2[periplasm] + a menaquinol[inner membrane] + H+[periplasm]
nitrate reduction III (dissimilatory)
nitrate + a menaquinol[inner membrane] + 2 H+
nitrite + a menaquinone[inner membrane] + H2O + 2 H+[periplasm]
nitrate reductase A, subunit
narH
dosCP Rho-independent terminator
dosC
diguanylate cyclase DosC
2 GTP
cyclic di-3',5'-guanylate + 2 diphosphate
diguanylate cyclase
dgcZ
rpoA
RNA polymerase, subunit
a nucleoside triphosphate + RNA(n)
RNA(n+1) + diphosphate
RNA polymerase, subunit
rpoB
yobF
cspA thermoregulator
apt
apt
fimA
fimAICDFGH
fimE
fimE
C0362
eamA
marC
yfhM
ydeA
idnR
cpxA
hscC
pepN
yhiS_1
Attenuator (Anti-terminator)
afuC
uxaA
ybbP
paaE
aroF-tyrA
yadV
fucA
pncB
nanCMS
ECK125135293
qseB
rapA
yciX
yeaR
rlmC
leuE
frlD
Attenuator (Terminator)
xylFGHR
xseA
gatZ
yjgN
wbbJ
yejM
insH21
ycjM
yneM
glyVXY
tgt
bluF
bcsB
iraM
cspC
yqiG
yqiH
ipeX
nupG
prlF
hfq
rnb
chbC
arrS
yebG
ltaE
pqqL
ariR
hrpA
ydiL
rpoD
lptD
mgrR
sroC
dusC
ydeE

Figure 2.5 *Network visualization of mutations in benzoate-adapted strains*

The shortest-paths network (234 nodes, 200 edges) generated by EVA analysis of benzoate-adapted strains. Nodes are highlighted by priority; high priority: red, unassigned: orange, low: green. The largest cluster in the shortest-paths network (125 nodes, 167 edges) contains the mutated transcription factor Rob and mutated the fabB -10 promoter element which have implications for fatty acid biosynthesis. Also represented in this cluster is the destruction of the aceBAK -10 promoter element in the E1-1 clade and a predicted damaging mutation in aceA in the C3-1 clade.

Discussion

Most existing methods for mutation analysis are limited to genes and require researchers to convert nucleotide variations to amino acid mutations. Additionally, mutations in a single strain are analyzed independently, making the investigation of mutation interactions a difficult task. Our approach accepts multiple formats and mutation information for one or more strains, automates a significant portion of analysis, and generates a network of mutations and their downstream biological features. Because direct mutation interactions and those occurring through regulation can be observed in EVA networks, our method can provide insight into underlying mechanisms affected by genomic mutations and support researchers in characterizing variant strains.

There remain many uncharacterized and poorly understood genomic features which could, in the future, be incorporated into the EVA pipeline. For example, when examining binding affinity of promoter sequences, we align a specific sequence to the lac promoter to use as a model. While the effect of gaps in the spacer region has been studied elsewhere, the energy matrix we employ as a scoring scheme does not address insertions and deletions. A meaningful penalty for gaps is not immediately clear as there is an absence of experimental data, however, this is not a limitation of the method presented by Kinney et al., and an expanded dataset could be included in the analysis. Strategies to analyze intergenic regions themselves, such as the distance between promoter elements could also be implemented in EVA.

Additionally, work has been done to characterize RBSs and RBS-promoter pairs in *E. coli* (Kosuri et al.; Na, Lee, & Lee; Salis, Mirsky, & Voigt) and effects on gene expression from 5-UTR and sRNA binding variants have been examined (Holmqvist, Reimegard, & Wagner). As more libraries of binding site variants are generated and associated mRNA and protein abundances quantified, more energy matrices will be available that will reveal how sequence variations affect phenotype. In the absence of binding site libraries and associated expression data, alignments of known binding site sequences to form positional weight matrixes could reveal acceptable variants similar to the method SIFT uses for coding regions. For TFBSs, motif data is retrieved from public databases (e.g., CollectTF (Kilic, White, Sagitova, Cornish, & Erill) and PRODORIC (Münch et al.)). Features such as Rho utilization sites are not currently available in public databases for *E. coli*, but as the Rho termination factor is believed to be responsible for terminating 20-50% of all mRNA synthesis in the organism (Koslover, Fazal, Mooney, Landick, & Block), this feature data would be a valuable addition to EVA. For mutated genes, mechanisms of transcription and translation efficiency, such as codon bias (Welch et al.) may help better understand silent mutations. Additionally, the Provean threshold could be recalibrated with the latest NCBI non-redundant database and specifically for bacteria.

In order to develop phenotype predictions, an expanded gene list including genes in the biological networks generated by EVA can be annotated with Gene Ontology (GO) terms (Ashburner et al.). This may be performed for all samples to search for overall evolutionary trends among biological replicates or for individual samples to examine a specific phenotype. While any single variation may be relevant to organism fitness, enriched GO terms from the expanded gene list can indicate importance to organism fitness and may capture biological knowledge of gene functions not represented in the EVA networks. However, as mutations in regulatory elements and genes that encode transcription factors will add additional, and sometimes functionally-related, genes into the EVA network, analysis for overrepresented GO terms could benefit from giving these downstream genes less weight.

One shortcoming of the per-mutation approach employed by EVA during the annotation step is the inability to interpret mutations in the context of one another. For example, a single genomic feature may contain multiple predicted mutations, but EVA would evaluate these separately. Future versions may consider such cases for improved interpretation. Another enhancement could include the integration of transcriptomic data or other omics data in the biological network that EVA generates.

Conclusions

EVA provides a framework to aid scientists in interpreting genetic variations that occur in metabolic evolution experiments by expanding annotations, prioritizing mutations. Indirect effects of mutations can be found in the biological network created by EVA that contains mutated features, downstream elements, and their interactions. It is important to note that we do not seek to quantify the effects of mutations but to offer a method of interpretation and constructive ranking to promote relevant laboratory experiments to further characterize mutation effects. EVA highlights the critical role of regulators and the need to include them in evolutionary experiment analyses. Software automation of mutation analysis in EVA improves upon what is generally a manual process. EVA is a principled approach to mutation analysis that can be refined as mechanisms of regulation are better understood and researchers perform more high-throughput and quantitative experiments to characterize regulatory sequences.

Acknowledgements

The authors would like to acknowledge the vision of Dr. Jacqueline Shanks for providing the impetus to systematically approach mutation analysis in microbial systems.

The authors would like to thank Cyverse, which is supported by the National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383, for computational resources.

Authors' contributions

EEB designed and implemented the EVA software and performed analysis of case study data. The project was conceived and advised by JAD and LRJ. All authors read and reviewed the final version of the manuscript.

References

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Research, 21(6), 961-973. doi:papers3://publication/uuid/D6083AF3-8E2E-4CF7-8AB1-004D9D58E3F8

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), 25-29. doi:papers3://publication/doi/10.1038/75556

Atsumi, S., Wu, T.-Y., Machado, I. M. P., Huang, W.-C., Chen, P.-Y., Pellegrini, M., & Liao, J. C. (2010). Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Molecular Systems Biology, 6(1), 449. doi:papers3://publication/uuid/E5E4CC66-CC2F-4343-8E7C-32F52AA61D28

Avihoo, A., & Barash, D. (2006). Shape similarity measures for the design of small RNA switches. Journal of biomolecular structure & dynamics, 24(1), 17-24. doi:papers3://publication/uuid/99472BFF-32B8-4A6C-94C5-C3095B4433F6

Bakke, I., Berg, L., Aune, T. E. V., Brautaset, T., Sletta, H., Tøndervik, A., & Valla, S. (2009). Random Mutagenesis of the Pm Promoter as a Powerful Strategy for Improvement of Recombinant-Gene Expression. Applied and Environmental Microbiology, 75(7), 2002-2011. doi:papers3://publication/uuid/625A2333-9944-4113-A3FA-24B5C5D5848F

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., . . . Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature, 461(7268), 1243-1247. doi:papers3://publication/uuid/3154160D-33D7-40D3-BFA6-4454AB6DB099

Benos, P. V. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic acids research, 30(20), 4442-4451. doi:papers3://publication/uuid/5B337CEB-08A9-46EB-96A4-2CC7A9A59807

Byrne, R. T., Klingele, A. J., Cabot, E. L., Schackwitz, W. S., Martin, J. A., Martin, J., . . . Cox, M. M. (2014). Evolution of extreme resistance to ionizing radiation via genetic adaptation of DNA repair. eLife, 3, e01322. doi:papers3://publication/uuid/AB20C77F-1350-4601-B79F-0877146AAC46

Castillo-Keller, M., Vuong, P., & Misra, R. (2006). Novel mechanism of Escherichia coli porin regulation. Journal of bacteriology, 188(2), 576-586. doi:papers3://publication/uuid/8A938060-CE01-4C53-8D00-32E83BC03A19

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino Acid substitutions and Indels. PLoS ONE, 7(10), e46688. doi:papers3://publication/uuid/53E1A263-4B6F-4386-A543-A5323B4F51BF

Cirz, R. T., Chin, J. K., Andes, D. R., de Crécy-Lagard, V., Craig, W. A., & Romesberg, F. E. (2005). Inhibition of Mutation and Combating the Evolution of Antibiotic Resistance. PLoS biology, 3(6), e176. doi:papers3://publication/uuid/85FE1D0E-4914-421D-B14C-D89D9C54DE13

Cordell, S. C., Robinson, E. J. H., & Lowe, J. (2003). Crystal structure of the SOS cell division inhibitor SulA and in complex with FtsZ. Proceedings of the National Academy of Sciences of the United States of America, 100(13), 7889-7894. doi:papers3://publication/uuid/863C434F-A6F2-4D62-9E8C-6B2CB8224F65

Creamer, K. E., Ditmars, F. S., Basting, P. J., Kunka, K. S., Hamdallah, I. N., Bush, S. P., . . . Slonczewski, J. L. (2017). Benzoate- and Salicylate-Tolerant Strains of Escherichia coli K-12 Lose Antibiotic Resistance during Laboratory Evolution. Applied and Environmental Microbiology, 83(2), e02736-02716. doi:papers3://publication/uuid/AE4D9EC2-2FAA-4F90-B7E5-068BDB7FC99C

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods in molecular biology (Clifton, N.J.), 1151, 165-188. doi:papers3://publication/uuid/68DB31A8-324E-4B2A-BE23-7988DC930923

den Dunnen, J. T., & Antonarakis, S. E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Human mutation, 15(1), 7-12. doi:papers3://publication/uuid/718A0FF6-2E04-454C-BFE7-5D4616F11BDD den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., . . . Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human mutation, 37(6), 564-569. doi:papers3://publication/uuid/C7D577F4-78B5-48BA-BDF4-741B9EA1FAC6

Finn, R. D. (2000). Escherichia coli RNA polymerase core and holoenzyme structures. The EMBO Journal, 19(24), 6833-6844. doi:papers3://publication/uuid/7BBA33D7-2AE2-47FD-BBC9-66ACDB96F409

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic acids research, 44(D1), D133-143. doi:papers3://publication/uuid/59309EE8-734B-4AAF-943A-0C4E0BA72C35

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D. C., & Shyr, Y. (2012). The effect of strand bias in Illumina short-read sequencing data. BMC genomics, 13(1), 666. doi:papers3://publication/uuid/735BC728-BB9F-4F16-A9B5-F86CE4C40850

Haft, R. J. F., Keating, D. H., Schwaegler, T., Schwalbach, M. S., Vinokur, J., Tremaine, M., . . . Landick, R. (2014). Correcting direct effects of ethanol on translation and transcription machinery confers ethanol tolerance in bacteria. Proceedings of the National Academy of Sciences, 111(25), E2576-2585. doi:papers3://publication/uuid/3E49031F-11C7-40DA-89A4-BC4A1F1329DD

Harden, M. M., He, A., Creamer, K., Clark, M. W., Hamdallah, I., Martinez, K. A., . . . Slonczewski, J. L. (2015). Acid-adapted strains of Escherichia coli K-12 obtained by experimental evolution. Applied and Environmental Microbiology, 81(6), 1932-1941. doi:papers3://publication/uuid/FD16E8AE-1E58-4F5A-BB32-38557CEBDEA3

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., . . . Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Molecular Systems Biology, 2, 2006.0007. doi:papers3://publication/uuid/894B5F1E-5948-4EE5-B8FD-1A31503782D6

Henikoff, J. G., & Henikoff, S. (1996). Blocks database and its applications. Methods in enzymology, 266, 88-105. doi:papers3://publication/uuid/83635ECC-9DEC-4F70-AA4E-3740E862D8EE

Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., . . . Palsson, B. Ø. (2006). Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nature genetics, 38(12), 1406-1412. doi:papers3://publication/uuid/48BBEEE0-A8DF-459F-B830-99E6003B7394

Holmqvist, E., Reimegard, J., & Wagner, E. G. H. (2013). Massive functional mapping of a 5'-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. Nucleic acids research, 41(12), e122-e122. doi:papers3://publication/uuid/DB9E1C8E-DBBC-49C7-A205-F6788C6B4720

Jang, Y.-S., Park, J. M., Choi, S., Choi, Y. J., Seung, D. Y., Cho, J. H., & Lee, S. Y. (2012). Engineering of microorganisms for the production of biofuels and perspectives based on systems metabolic engineering approaches. Biotechnology advances, 30(5), 989-1000. doi:papers3://publication/uuid/FB1A1F51-5EC5-4634-A856-005C12A0F9A4

Janion, C. (2008). Inducible SOS Response System of DNA Repair and Mutagenesis in Escherichia coli. Int J Biol Sci, 338-344. doi:papers3://publication/uuid/E7A6C3C4-F046-4F07-87F5-FD2A4924F9E1

Jin, T., Chen, Y., & Jarboe, L. R. (2016). Evolutionary Methods for Improving the Production of Biorenewable Fuels and Chemicals: Biotechnology for Biofuel Production and Optimization.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., . . . Karp, P. D. (2013). EcoCyc: fusing model organism databases with systems biology. Nucleic acids research, 41(Database issue), D605-612. doi:papers3://publication/doi/10.1093/nar/gks1027

Kilic, S., White, E. R., Sagitova, D. M., Cornish, J. P., & Erill, I. (2013). CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic acids research, 42(D1), D156-D160. doi:papers3://publication/uuid/017A868B-63A6-455F-97D3-463C0C9024C6

Kinney, J. B., Murugan, A., Callan, C. G., & Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences, 107(20), 9158-9163. doi:papers3://publication/uuid/10B7995D-62E7-41A1-9349-3F3E7417C2A4

Kiryu, H., & Asai, K. (2012). Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations. Bioinformatics, 28(8), 1093-1101. doi:papers3://publication/uuid/A517FA08-F975-4618-94FE-AC7E6109A9CB

Koslover, D. J., Fazal, F. M., Mooney, R. A., Landick, R., & Block, S. M. (2012). Binding and translocation of termination factor rho studied at the single-molecule level. Journal of Molecular Biology, 423(5), 664-676. doi:papers3://publication/uuid/541D4917-FDBF-4A65-86F4-E4132A70F724

Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., . . . Church, G. M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proceedings of the National Academy of Sciences, 110(34), 14024-14029. doi:papers3://publication/uuid/2D59C572-41AB-4F2F-8458-FA8B0941567D

LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., . . . Feist, A. M. (2015). Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Applied and Environmental Microbiology, 81(1), 17-30. doi:papers3://publication/doi/10.1128/AEM.02246-14

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359. doi:papers3://publication/doi/10.1038/nmeth.1923

Lessard, I. A., & Walsh, C. T. (1999). VanX, a bacterial D-alanyl-D-alanine dipeptidase: resistance, immunity, or survival function? Proceedings of the National Academy of Sciences of the United States of America, 96(20), 11028-11032. doi:papers3://publication/uuid/D154A322-9DDD-437A-96B5-01ABCA6A830F

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady. doi:papers3://publication/uuid/72AB0F7F-E180-4E00-9C6D-DE9E44CE9796

Malhotra, A., Severinova, E., & Darst, S. A. (1996). Crystal structure of a sigma 70 subunit fragment from E. coli RNA polymerase. Cell, 87(1), 127-136. doi:papers3://publication/uuid/DB98376A-4C31-4251-A74E-0C0F1E5F7A62

McKenzie, G. J., Harris, R. S., Lee, P. L., & Rosenberg, S. M. (2000). The SOS response regulates adaptive mutation. Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6646-6651. doi:papers3://publication/uuid/27D3D29A-796D-41FA-B0C7-18AC0E0322DB

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS biology, 14(1), e1002342. doi:papers3://publication/uuid/7FD17A17-0881-4CE5-9E86-5F644A479FCF

Møller, T. S. B., Rau, M. H., Bonde, C. S., Sommer, M. O. A., Guardabassi, L., & Olsen, J. E. (2016). Adaptive responses to cefotaxime treatment in ESBL-producing Escherichia coli and the possible use of significantly regulated pathways as novel secondary targets. Journal of Antimicrobial Chemotherapy, 71(9), 2449-2459. doi:papers3://publication/uuid/CC55EC01-8447-4476-8A77-F244F1AE5E1E

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., & Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. Nucleic acids research, 31(1), 266-269. doi:papers3://publication/uuid/3905D60C-1A64-4AF5-8EB2-5CE31799F827

Na, D., Lee, S., & Lee, D. (2010). Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. BMC systems biology, 4, 71. doi:papers3://publication/uuid/C368407F-9A15-4021-B37F-C3BE16900588

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Research, 11(5), 863-874. doi:papers3://publication/uuid/EC62C41C-EBF6-48F6-9240-41555CF4292F

Reyes, L. H., Winkler, J., & Kao, K. C. (2012). Visualizing evolution in real-time method for strain engineering. Frontiers in microbiology, 3, 198. doi:papers3://publication/uuid/24D5FEC6-C8BA-47FD-A828-FECD00F00E89

Riley, M. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic acids research, 34(1), 1-9. doi:papers3://publication/uuid/4C8C7046-3F14-4548-A27E-6D54081DB321

Sabarinathan, R., Tafer, H., Seemann, S. E., Hofacker, I. L., Stadler, P. F., & Gorodkin, J. (2013). RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. Human mutation, 34(6), 925-925. doi:papers3://publication/uuid/4E84C142-5B9F-4654-93CB-828641FDB696

Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nature biotechnology, 27(10), 946-950. doi:papers3://publication/uuid/B113D5EA-F3DD-4EFF-BFBB-A6945A261E80

Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., . . . Feist, A. M. (2014). Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Molecular biology and evolution, 31(10), 2647-2662. doi:papers3://publication/uuid/24172F26-88F2-4322-8996-36A116AE9C52

Shannon, P. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 13(11), 2498-2504. doi:papers3://publication/uuid/2FBEA35A-626F-4C7B-901B-AFC9F9E99568

Stormo, G. D., & Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nature Publishing Group, 11(11), 751-760. doi:papers3://publication/uuid/40AFC6FD-ACBB-4BD5-A7BE-6652DC10CA9D

Utrilla, J., Licona-Cassani, C., Marcellin, E., Gosset, G., Nielsen, L. K., & Martinez, A. (2012). Engineering and adaptive evolution of Escherichia coli for d-lactate fermentation reveals GatC as a xylose transporter. Metabolic Engineering, 14(5), 469-476. doi:papers3://publication/uuid/35A53062-2D32-4B85-9DAC-B896C606C17B

Wang, K. K., Stone, L. K., Lieberman, T. D., Shavit, M., Baasov, T., & Kishony, R. (2016). A Hybrid Drug Limits Resistance by Evading the Action of the Multiple Antibiotic Resistance Pathway. Molecular biology and evolution, 33(2), 492-500. doi:papers3://publication/uuid/3A3A781D-6F61-4E28-9428-37D55CCCAF58

Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J., & Gustafsson, C. (2009). Design parameters to control synthetic gene expression in Escherichia coli. PLoS ONE, 4(9), e7002. doi:papers3://publication/uuid/B66B3CDB-9CC4-4971-B477-7748DE374502

Worth, C. L., Preissner, R., & Blundell, T. L. (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic acids research, 39(Web Server issue), W215-222. doi:papers3://publication/uuid/957C9048-3923-4221-B82F-43C874D282F4

Yu, W., Bing, L., & Zhenhua, L. (2009). AmpC Promoter and Attenuator Mutations Affect Function of Three Escherichia coli Strains. Current Microbiology, 59(3), 244-247. doi:papers3://publication/uuid/2E736F63-A6F4-4EDF-9ECD-853C6E212952

CHAPTER 3. GENOME-LEVEL REVERSE ENGINEERING OF *ESCHERICHIA COLI EVOLVED* FOR INCREASED SHORT-CHAIN FATTY ACID TOLERANCE AND PRODUCTION

A research paper to be submitted to Metabolic Engineering

Yingxi Chen[1], Erin E. Boggess[2], Julie A. Dickerson[2], Laura R. Jarboe[1]

[1]*Department of Chemical & Biological Engineering, Iowa State University, Ames, IA, 50011.*

[2]*Department of Electrical & Computer Engineering, Iowa State University, Ames, IA, 50011.*

Abstract

Metabolic evolution is a valuable strategy for overcoming toxicity of target biorenwable chemicals, however reverse engineering of evolved strains is necessary to understand how the tolerance phenotype is achieved. Whole genome sequencing and mutation analysis are required to identify the genomic changes that occur during an evolution experiment and characterization of both single and multiple mutations is necessary to understand their individual and combined contributions to phenotype. Here, we analyze the genome of *Escherichia coli* evolved for improved exogenous octanoic acid tolerance and reconstruct mutations in the parent strain to determine their contribution to phenotype. We identified mutations in *rpoC*, *basR*, and *basS* in strains evolved for tolerance and an insertion sequence in *waaG* in the parent strain that was lost during the course of the experiment, restoring function. We find the repair of *waaG* to reduce the amount of extracellular polysaccharides produced by the cells as well as decrease leakage and improve the specific growth rate in an octanoic acid challenge experiment. The *rpoC* mutation further improves tolerance after *waaG* is repaired and the mutations in *basS* and *basR* are found to improve

cell membrane integrity. These results highlight strategies to overcome membrane damage as a result of octanoic acid toxicity as well as the importance of studying synergistic effects among mutations found in evolved strains.

Introduction

Fatty acids are of great importance in the industrial field (Jarboe, Royce, & Liu, 2013) due to their wide applications as multifunctional precursors to produce various fuels, chemicals, and textile fibers (Dellomonaco, Fava, & Gonzalez, 2010; Perez, Richter, Loftus, & Angenent, 2013; Zhang, Yang, Yang, & Ma, 2009). So far, the production of industrial fatty acids relies heavily on a nonrenewable and unsustainable resource, petroleum (Dellomonaco et al., 2010; C. Zhang et al., 2009), which can lead to severe environmental, political, and economic consequences (Stephanopoulos, 2007). Therefore, it is necessary to develop new pathways to produce fatty acids using renewable and sustainable carbon feedstocks. In this respect, biocatalysts are attractive and promising. They have already been broadly applied to biorenewable industries for the production of various chemicals, such as ethanol, glycerol, 1, 3-propanediol, and lactic acid (Nikolau, Perera, Brachova, & Shanks, 2008). Moreover, it is potentially feasible for researchers to engineer organisms to obtain target biocatalysts with significant ability to produce fatty acids by utilizing microorganisms that can naturally synthesize fatty acids with 12-18 carbons, the primary components of the cell membrane (Nikolau et al., 2008). So far, remarkable progress has been achieved by researchers for the realization of the production of fatty acids by biocatalysts on a commercial level (Jarboe, Liu, & Royce, 2011; Lennen & Pfleger, 2012; L. A. Royce et al., 2014; L. A. Royce, Liu, Stebbins, Hanson, & Jarboe, 2013; Volker et al., 2014; Wu, Karanjikar, & San, 2014; Wu, Lee, Karanjikar, & San, 2014).

Researchers have found that fatty acids produced by microbes are toxic to the microbes themselves at concentrations below the desired yield and titer (Jarboe et al., 2013; Lennen et al., 2011; Volker et al., 2014), just like other attractive biofuels and biorenewable chemicals (Baer, Blaschek, & Smith, 1987; Huffer, Roche, Blanch, & Clark, 2012; Yomano, York, & Ingram, 1998). This is a major obstacle for boosting the yield and titer of fatty acids (Jarboe et al., 2013). The mechanism of the toxicity of fatty acid to *Escherichia coli* has been studied, and it was reported that fatty acids can lower the cell viability by damaging the cell membrane and decreasing intracellular pH (Jarboe et al., 2013; Lennen et al., 2011; L. A. Royce et al., 2014; L. A. Royce et al., 2013). Several groups have tried to overcome fatty acid toxicity in *E. coli* by employing different approaches. One strategy was to restore the cell membrane to improve cell viability by over-expressing the fatty acid biosynthesis regulator, *fabR*, and introducing two foreign acyl-ACP thioesterase genes in *E. coli*. However, the engineered *E. coli* did not show improved fatty acid productivity (Lennen et al., 2011; Lennen & Pfleger, 2013). Another approach was to delete the acyl-ACP synthase (*aas*) gene in *E. coli*, which resulted a decreased percentage of medium chain fatty acids in the membrane, increased tolerance to medium chain fatty acids, and a slightly improved yield of fatty acids (Sherkhanov, Korman, & Bowie, 2014). Finally, metabolic evolution was employed as a strategy to increase tolerance to short chain fatty acids (SCFAs) (L. A. Royce et al., 2013). The evolved strain resulting from the short-term adaptation experiment exhibits an increased SCFA tolerance phenotype and improved production titer of the SCFAs (L. A. Royce et al., 2015).

Reverse engineering aims to both identify mutations that contribute to the phenotype of evolved strains and understand why these mutations are beneficial. It is a powerful tool to elucidate the mechanisms behind the evolution experiments, which can be used as design strategies for improving tolerance and production in other engineered strains. Identifying and understanding the key mutations that support the evolved phenotype requires knowledge of what mutations occurred during the evolutionary engineering. Whole-genome sequencing is invaluable for finding genomic mutations (L. Royce, Boggess, Jin, Dickerson, & Jarboe, 2013) and transcriptome analysis and metabolic flux analysis have been proven useful for revealing the underlying mechanisms of the mutations (Elliot N. Miller et al., 2009; E. N. Miller et al., 2009). After mutations are found, the next steps are to explore which mutations promote fitness, the mechanisms of how the tolerance to inhibitors has increased, and the functions of poorly-characterized enzymes and pathways involved in the evolved phenotype. For instance, in an isobutanol tolerance study, five mutations were identified as primarily responsible for increased tolerance, and glucosamine-6-phosphate was identified as an important metabolite for isobutanol tolerance in *E. coli* (Atsumi et al., 2010). Increasing furfural tolerance was achieved by silencing the NADPH-dependent oxidoreductase gene (*yqhD* and *dkgA*) in *E. coli* (E. N. Miller et al., 2009). The new glucose uptake system and mechanism of increased ATP level in the evolved strain has been well studied, which were the key mechanisms of improving succinate production in *E. coli* (X. Zhang et al., 2009).

In this work, we applied reverse engineering to study *E. coli* strains evolved for increased octanoic acid (C8) tolerance. To understand the genotype-phenotype relationship, the whole-genome sequencing of the evolved and parent strains was performed. The parent strain, ML115, is a MG1655 derivative by knocking out three genes and adding an antibiotic marker (Δ*fadD,* Δ*poxB,* and Δ*ackA-pta:cm*^{R}) in order to inactivate the fatty acid beta-oxidation pathway and two acetate production pathways. The order in which mutations were acquired was also determined in this work. Reconstructed strains with both single and multiple mutations were used in phenotypic characterization experiments in order to identify individual and combined contributions to fitness.

Materials and methods

Strains, plasmids and bacterial cultivation

All bacterial strains and plasmids used in this study are listed in Table 3.1 and Table 3.2. *E. coli* DH5α was used as a cloning strain, while the parent strain *E. coli* ML115, and the evolved strain LAR1 were used in the genome modification procedures. All *E. coli* strains were grown overnight at 37°C with 250 rpm orbital shanking in 25 mL of MOPS minimal media (Wilmes-Riesenberg & Wanner, 1992) with 2.0% (w/v) glucose and chloramphenicol (35 µg/mL, if needed) in 250 mL baffled flasks. The overnight cultures were diluted to 0.05 of optical density at 550 nm (OD_{550}) for the octanoic acid tolerance test or diluted to 0.1 at OD_{550} for testing membrane leakage, membrane fluidity, cell hydrophobicity, and cell membrane composition. *E. coli* transformants were grown in media at 37°C, or 30°C, with chloramphenicol (35 mg/L), ampicillin (100 mg/L), kanamycin (50 mg/L), or spectinomycin (50 mg/L) as needed.

Table 3.1 *Plasmids used in this study.*

Plasmids	**Description**	**Reference**
pKD4	FRT-Kan-FRT cassette template, Amp^{R}, Km^{R}	(Datsenko & Wanner, 2000)
pKD46	λ Red recombinase expression plasmid, Amp^{r}	(Datsenko & Wanner, 2000)
pCP20	FLP recombinase expression, Amp^{R}, Cm^{R}	(Datsenko & Wanner, 2000)
pUC57-rpoC-A	rpoC-A-FRT-Kan-FRT cassette template, Km^{R}	This study
pUC57-rpoC-C	rpoC-C-FRT-Kan-FRE cassette template, Km^{R}	This study

Table 3.2 *Strains used in this study.*

All strains contain the Cm^R chloramphenicol resistance gene.

Strain	**Mutations**				**Editing Method**	**Reference**
	***waaG* IS removed**	***rpoC***	***basR***	***basS***		
ML115						(L. A. Royce et al., 2015)
LAR1	*	*	*			(L. A. Royce et al., 2015)
LAR2	*	*		*		(L. A. Royce et al., 2015)
ML115+waaGInD	*				CRISPR/Cas9	This study
ML115+rpoC*		*			Lambda Red	This study
ML115+basR*			*		CRISPR/Cas9	This study
ML115+basS*				*	Lambda Red	This study
ML115+waaGInD+rpoC*	*	*			CRISPR/Cas9	This study
ML115+waaGInD+basR*	*		*		CRISPR/Cas9	This study
ML115+waaGInD+basS*	*			*	CRISPR/Cas9	This study
ML115+rpoC*+basR*		*	*		CRISPR/Cas9	This study
ML115+basR*+basS*			*	*	CRISPR/Cas9	This study
ML115+waaGInD+rpoC*+basR*	*	*	*		CRISPR/Cas9	This study
ML115+waaGInD+basR*+basS*	*		*	*	CRISPR/Cas9	This study
LAR1+rpoC	*		*		CRISPR/Cas9	This study

Whole-genome sequencing and mutation verification

The genomic DNA of ML115, LAR1 and LAR2 was purified using the Qiagen Blood and Tissue kit. The Illumina Genome Analyzer II platform was used for high throughput sequencing with 77 bp (base pair) paired-end reads as described (L. Royce et al., 2013). All samples were run on a single lane. Breseq version 0.31.0, a pipeline for finding mutations in microbial genomes, was used to align short read data and predict mutations (Deatherage & Barrick, 2014). Bowtie2 version 2.3.3 (Langmead & Salzberg, 2012) and R version 3.4.1 (R Core Team, 2018) software were used in the breseq pipeline. The U00096.3 genome for *E. coli* K-12 MG1655 (Blattner et al., 1997; Hayashi et al., 2006) was used as the reference sequence to which short-read data from both parent and evolved strains were aligned. 87.8,%, 91.0%, and 87.3% of reads were successfully aligned to the reference sequence for ML115, LAR1, and LAR2, respectively. Previous genomic interventions (Δ*fadD,* Δ*poxB,* and Δ*ackA-pta:cm*^{R}) present in ML115 were verified as regions of missing coverage.

When considering predicted mutations, we followed the filters recommended by breseq to reduce the number of false positives. A Fisher's exact test was performed for the distribution of reads aligning in the forward and reverse direction for the reference and variant sequence. If the distribution skewed to favor alignment in one direction, this may indicate a sequencing error in reads on one strand. Additionally, a Kolmogorov-Smirnov test was performed to test if the base quality scores corresponding to variant sequences are lower than the quality scores corresponding to the reference sequence.

Genomic variations displaying neither strand bias nor lower quality scores compared to the reference sequence which are predicted in an evolved strain but not the parent strain (and vice versa) were selected as mutations of interest to be verified with polymerase chain reaction (PCR) and Sanger sequencing.

Genes containing predicted genomic variations along with an additional 500 bp upstream and downstream of the coding region were sequenced in order to verify mutations. Target gene fragments were PCR amplified with Qiagen Taq PCR master mix, primers, and the genome of evolved strain was used as the template. All primers were designed by Primer3 software (Untergasser et al., 2012) and synthesized by Integrated DNA Technologies (IDT). The sizes of PCR products were initially examined on a 1% TAE agarose gel with a 1 Kb plus DNA ladder. Next, PCR products exhibiting the expected gene fragment size were purified by QIAquick PCR purification kits (Qiagen) and submitted to Iowa State University DNA facility for DNA sequencing. The sequencing results were aligned to the *E. coli* K-12 MG1655 genome using the online NCBI standard nucleotide BLAST software (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to confirm the mutations. Mutations were verified by repeating all steps above using the genome of evolved strain and parent strain separately as templates.

PCR-restriction fragment length polymorphism (PCR-FRLP)

We applied RCR-FRLP to determine the order of mutations in evolved strains, which use cells culture saved after every transfer during adaptive evolution as DNA template. For the mutation in *rpoC* (A1256C), the 660 bp DNA fragment which includes the mutant point was amplified by PCR with the primers rpoCCF, rpoCCR, and DreamTaq Green PCR master mix 2X (Thermo Fisher Scientific). The PCR products were purified by DNA Clean &

Concentration kit (Zymo). Approximately 10 µl of purified PCR products were digested with restriction enzyme BsaJI (New England Biolabs) according to the manufacturer's instruction. The restriction fragments were separated on a 1% TAE agarose gel with 1 Kb plus DNA ladder. The pair of primers basRCF and basRCR, and restriction enzyme SfcI were used for *basR* mutation. The basSCF and basSCR primers, and restriction enzyme FatI were used for *basS* mutation. For the *waaG* mutation, only waaGCF and waaGCR primers were needed.

Genomic manipulations

All genomic manipulations were carried out using either lambda red recombinase system (Datsenko & Wanner, 2000) or CRISP-cas9 system (Jiang et al., 2015). For the lambda red recombinase system, *E. coli* strains were first transformed by electroporation to harvest the pKD46 plasmid, and then the lambda red recombinases were induced by adding L-arabinose (2 mM). The kanamycin resistance cassette was amplified from plasmid pKD4 by PCR using primers with flanking homologous regions for the target gene, except the rpoC(1256A)+kan, and rpoC(1256C)+kan cassettes which were synthesized by GenScript company. The purified PCR products were transformed into the electro-competent *E. coli* cells harboring pKD46, and lambda red recombinases system was induced. The resulting kanamycin resistant colonies were screened for the successful gene replacement by the PCR amplification, restriction enzyme digestion, and DNA sequencing. The scarless CRISPR-Cas9 approach was also applied to achieve gene replacement in parent and evolved strains.

Octanoic acid tolerance test

Octanoic acid tolerance was determined by measuring OD_{550} every hour. Overnight seed cultures were inoculated into 250 mL baffled flasks, which contained 25 mL MOPS with 2.0% (w/v) dextrose and 10 mM octanoic acid (1.44 g/L) with an initial media pH of 7.0 and an initial OD_{550} of 0.05. The control groups used the same culture without the addition of octanoic acid. The flasks were incubated in a rotary shaker at 200 rpm and 37°C. Cultures were taken and measured at OD_{550} every hour.

Determination of fatty acid titers

Fatty acid production was quantified by measuring total fatty acids via an Agilent 6890 gas chromatograph coupled to an Agilent 5973 mass spectroscope (GC-MS) after fatty acid extraction and derivatization (Torella et al., 2013). Briefly, fatty acid extraction was done as follows: 1 mL culture was transferred into a 2 mL microcentrifuge tube, 125 µL 10% NaCl (w/v), 125 µL acetic acid, 20 µL internal standard (1 µg/µL C7, C11, C15 in ethanol), 500 µL Ethyl Acetate was added subsequently. The mixture was vortexed 30 seconds and centrifuged at 16000 g for 10 minutes, then the 250 µL top layer, which contained free fatty acid, was transferred into a glass tube. For the fatty acid derivatization part, 2.25 mL 30:1 EtOH: 37% HCl (v/v) was added into the glass tube from fatty acid extraction part, incubated at 55°C for 1 hour, then cooled to room temperature. After this, 1.25 mL ddH_2O and 1.25 ml hexane was added followed by vortexed and centrifuged at 2,000 × g for 2 minutes. The top hexane layer was then analyzed by GC-MS using the following programs: the initial temperature was set at 50°C, holding for 1 minute, with the following temperature ramp: 20°C/minute to 140°C, 4°C/minute to 220°C, and 5°C/minute to 280°C with 1 ml/minute helium carrier gas. The relative retention factor of C7/C11/C15 was used to adjust the relative amounts of the individual fatty acids analyzed.

Membrane characterization

Membrane permeability

The seed culture was inoculated into 250 mL baffled flasks with 25 mL MOPS media with 2.0% (w/v) glucose. The flasks were incubated in a rotary shaker at 250 rpm at 37°C. Cells were then harvested at mid-log phase (OD_{550} ≈ 1) followed by centrifugation at 4,500 × g and 22°C for 10 minutes. The cells were then treated with PBS with 10 mM octanoic acid at pH 7.0, incubated at 37°C for 1 hour along with a control group to which no octanoic acid was added. Subsequently, cells were centrifuged at 4,500 × g at 4°C for 10 minutes, washed twice with PBS (pH 7.0), and resuspended in PBS at a final OD_{550} = 1. Then, 100 µL resuspended cell solution was diluted with 900 µL PBS. Cells with damaged membrane were stained by the addition of 1 µL of 5 mM SYTOX green in dimethyl sulfoxide (Invitrogen, Carlsbad, CA), and tested by flow cytometric analysis performed on the BD Biosciences FACSCanto II (Lennen & Pfleger, 2013). Approximately 18,000 events were tested per sample, and each sample had three parallel groups.

Membrane fluidity

Membrane fluidity was tested using 1, 6-diphenyl-1, 3, 5-hexatriene as previously described (Mykytczuk, Trevors, Leduc, & Ferroni, 2007; L. A. Royce et al., 2013). Briefly, *E. coli* cells were treated as described in membrane permeability section. Cell pellets were re-suspended in PBS at a final OD_{550} = 1, then 500 µL resuspended cell solution was transferred into a 1.5 mL centrifuge tube which contained 500 µL 0.4 µM DPH, vortexed, and incubated at 37°C. Samples were then centrifuged and cell pellets were resuspended with 500 µL PBS (pH 7.0). 100 µL cell solution was transferred into sterile black-bottom Nunclon delta surface

96-well plate with 4 replicates, and the cell solution without DPH was used as blank. Membrane fluorescence polarization was measured using a Synergy 2 Multi-Mode microplate reader from BioTek.

Cell surface hydrophobicity

Cells were treated with MOPS with 2.0% (w/v) glucose and 10 mM octanoic acid at pH 7.0, incubated at 37°C for 1 hour along with a control group to which no octanoic acid was added. Un-adapted and 10 mM C8 adapted cells were washed twice with PBS, and resuspended in PBS (pH 7.0) to $OD_{550} \approx 0.6$. Then, 4 mL cells were added to a glass tube and 100 μL resuspended cells were used to measure $OD_{550,}$ recorded as OD_1. Next, 1 mL dodecane was added to the glass tubes (Pembrey, Marshall, & Schneider, 1999). The glass tube was vortexed using a multi-tube vortexer (Thermo Fisher Scientific Inc., Waltham, MA, USA) at 2500 rpm for 10 minutes to homogenize the aqueous and organic phases. The glass tube was left to stand for 15 minutes to allow phase separation and the OD_{550} of the aqueous phase (OD_2) was determined. Partitioning of the bacteria suspension is calculated using the following equation:

$$Percent\ partitioning = \frac{OD_1 - OD_2\ of\ aqueous\ phase}{OD_1} * 100$$

Membrane lipid composition

E. coli cells were harvested at mid-log phase, resuspended in 25 mL MOPS media with 2.0% (w/v) dextrose and 30 mM C8 at pH 7.0 and incubated for 3 hours at 37°C. The cells were washed twice in cold sterilized water and resuspended into 6 mL methanol. 1.4 mL cells solution was transferred into glass tubes with three replicates (Bligh & Dyer, 1959), and sonicated for three, 30 second bursts. A total of 20 μL of 1 μg/ μL C7, C11, and C15 in methanol was added as an internal standard. For fatty acid extraction, the glass tube was incubated at 70°C for 15 minutes and cooled to room temperature. The cells were centrifuged at 4000 rpm for 5 minutes. The supernatant was transferred into a new glass tube with 1.4 mL nanopure water, and the mixture was vortexed. Chloroform with a volume of 750 μL was add into the cell pellets, vortexed and shaken in a horizontal shaker at 150 rpm, 37°C for 5 minutes. We transferred the supernatant with H_2O back to the cell pellets glass tube, vortexed for 2 minutes, then centrifuged at 3,000 rpm for 5 minutes. The lower chloroform layer which contain free fatty acid was transferred into a new glass tube. The free fatty acids were concentrated with an N-Evap nitrogen tree evaporator. For fatty acid derivatization, 2 mL of 1N HCl was added in methanol to the samples. The free fatty acids were concentrated under nitrogen, heated to 80°C for 30 minutes, and then cooled to room temperature. 2 mL of 0.9% NaCl solution and 1 mL hexane was added and followed by vortex for 2 minutes and centrifugation at 2,000 × g for 2 minutes. The upper layer, the hexane with FAMEs, was transferred into a GC vial for analysis. The GC-MS was equipped with the same instruments as that used in the determination of fatty acid titers. The ratio of saturated to unsaturated fatty acids (S:U) and weight-average lipid length were calculated as previously described (L. A. Royce et al., 2013).

Results and Discussion

Verified mutations in evolved strains

To identify mutations acquired during the metabolic evolution experiment, we sequenced the genomic DNA of ML115, LAR1, and LAR2 using the Illumina platform. We used the breseq pipeline and short-read aligner, Bowtie2, to map reads from each strain to the *E. coli* K-12 MG1655 reference genome and identify sequence variations. A 768 bp insertion sequence (IS) was predicted in *waaG* in the parent strain, but neither of the evolved strains. The same mutation in *rpoC* was predicted in both evolved strains and results in an amino acid change from histidine to proline at position 419 in RpoC. Each evolved strain exhibits a mutation related to the BasS-BasR two-component signal transduction system. In LAR1, *basR* has a point mutation that results in an amino acid change from aspartic acid to tyrosine at position 28 in the protein product. In LAR2, *basS* has a 27 bp deletion that results in a 9 amino acid (aa) deletion in BasS. Mutations predicted in both the parent and evolved strains were not considered for further analysis. Computationally predicted mutations were verified by PCR and Sanger sequencing.

Genome diff files from the breseq output for the parent and evolved strains were submitted as input to the EVA pipeline (Boggess, Jarboe, & Dickerson, 2018) for additional analysis. HGVS-style descriptions (den Dunnen et al., 2016) of amino acid variations were generated by the EVA pipeline where applicable. Provean scores (Choi, Sims, Murphy, Miller, & Chan, 2012), which provide an indication of whether a mutation may be damaging (score ≤ -2.5) or tolerated (score > -2.5) and EVA prioritization of mutations are provided in Table 3.3.

EVA analysis also generates a network representation of mutated features and downstream biological features that may be influenced by the mutation through gene regulation and cellular metabolism (Figure 3.1). However, we find that the latest release of RegulonDB, version 9.4 (Gama-Castro et al., 2016), does not include some published regulatory activities for the BasR transcription factor, specifically transcription activation of *waaH* (computationally predicted), *eptA*, and *arnBCADTEF* (Froelich, Tran, & Wall, 2006; H. Ogasawara, S. Shinohara, K. Yamamoto, & A. Ishihama, 2012). We manually supplemented EVA with nine transcription activation links from BasR to each of these genes. From this updated visualization, we do not find any immediate relationship among mutations that occur in any individual strain. However, the relationship between *basS* and *basR* is clear in the network and genes in the BasR regulon are candidates for further study as their expression may be affected by either mutation in the evolved strains.

Pathways included in the EVA-generated network include proline degradation and proline to cytochrome electron transfer, the QseBC quorum-sensing two-component system, which is involved in regulation of flagella biosynthesis, and polymyxin resistance. In addition, many genes in the BasR regulon are located in the membrane: *putA*, *eptA*, *dgkA*, *waaH*, *qseC*, *arnCDTEF*, and *csgDEFG*. Some of the genes in the BasR regulon, such as *qseB*, *putA*, and *csgD* also encode transcription factors which may alter expression of additional genes through transcription regulation. The *eptA* and *waaH* genes are particularly interesting because of their role in modifying LPS as is the *csg* operon for its relevance to curli assembly and biofilm formation.

It must also be noted that interactions between RNA polymerase and promoter sequences are numerous and not represented in this network, but a query of the RegulonDB database identifies 1,606 genes with associated sigma 70 promoters. Because the *rpoC* mutation could affect transcription initiation of a large number genes, additional analysis may benefit from a transcriptomic experiment.

Table 3.3 *EVA annotation and prioritization of mutations in ML115, LAR1, and LAR2.*

Mutations are ordered by position in the genome. IS indicates the introduction of an insertion sequence at the specified position.

ML115	LAR1	LAR2	Position	Annotation	b-number	Mutation	HGVS description	Provean score	EVA priority
*	*	*	144,786	*yadI*	b0129	G → T	A70A	N/A	Low
*	*	*	1,704,001	*ydgJ*	b1624	A → C	Q103P	-3.86	High
*	*	*	1,873,031	*dgcJ*	b1786	IS	R331_S496delinsGCTSVYTKMCREKILVMR	-387.40	High
*	*	*	1,946,308	*yebB*	b1862	IS	G20_V200delinsVLPYLVKYQLHQIAGVITSGSLSVITVKTSWLQKAGFPFQPSPRYLVLLNVRLINAML	-451.03	High
*	*	*	2,610,245	*hyfH*	b2488	G → A	G28S	-2.61	High
*			3,806,607	*waaG*	b3631	IS	H154_G374delinsLIKLNLNVFKFFLPVFIRTENTVSKSQTAVKFIARKMA	-674.93	High
	*	*	4,186,605	*rpoC*	b3899	A → C	H419P	-9.35	High
*	*	*	4,296,381	intergenic (*gltP*/*yjcO*)	(b4077/b4078)	+GC		N/A	Unassigned
		*	4,332,397	*basS*	b4112	Δ27 bp	A285_G293del	-25.51	High
	*		4,333,869	*basR*	b4113	C → A	D28Y	-7.76	High

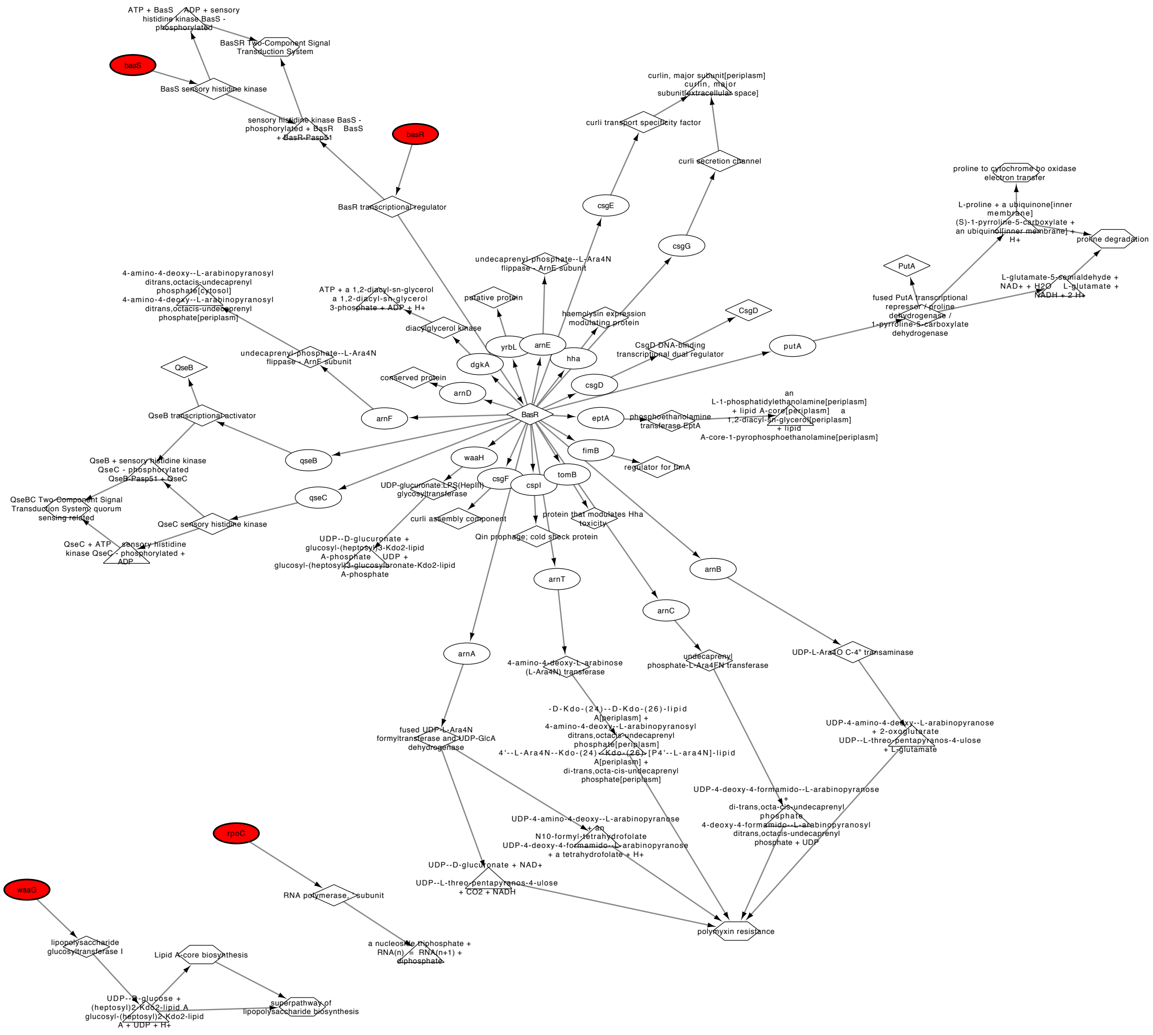

ATP + BasS ADP + sensory histidine kinase BasS - phosphorylated
BasSR Two-Component Signal Transduction System
basS
BasS sensory histidine kinase
sensory histidine kinase BasS - phosphorylated + BasR BasS + BasR-Pasp51
basR
BasR transcriptional regulator
curlin, major subunit[periplasm] curlin, major subunit[extracellular space]
curli transport specificity factor
curli secretion channel
csgE
csgG
proline to cytochrome bo oxidase electron transfer
L-proline + a ubiquinone[inner membrane] (S)-1-pyrroline-5-carboxylate + an ubiquinol[inner membrane] + H+
proline degradation
PutA
fused PutA transcriptional repressor / proline dehydrogenase / 1-pyrroline-5-carboxylate dehydrogenase
L-glutamate-5-semialdehyde + NAD+ + H2O L-glutamate + NADH + 2 H+
undecaprenyl-phosphate--L-Ara4N flippase - ArnE subunit
4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[cytosol] 4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[periplasm]
ATP + a 1,2-diacyl-sn-glycerol a 1,2-diacyl-sn-glycerol 3-phosphate + ADP + H+
putative protein
diacylglycerol kinase
haemolysin expression modulating protein
CsgD
CsgD DNA-binding transcriptional dual regulator
putA
undecaprenyl-phosphate--L-Ara4N flippase - ArnF subunit
yrbL
arnE
hha
dgkA
QseB
conserved protein
arnD
csgD
QseB transcriptional activator
arnF
BasR
eptA
an L-1-phosphatidylethanolamine[periplasm] + lipid A-core[periplasm] a 1,2-diacyl-sn-glycerol[periplasm] + lipid A-core-1-pyrophosphoethanolamine[periplasm]
phosphoethanolamine transferase EptA
qseB
fimB
regulator for fimA
QseB + sensory histidine kinase QseC - phosphorylated QseB-Pasp51 + QseC
waaH
csgF
tomB
cspI
qseC
QseBC Two-Component Signal Transduction System: quorum sensing related
UDP-glucuronate:LPS(HepIII) glycosyltransferase
QseC sensory histidine kinase
curli assembly component
protein that modulates Hha toxicity
QseC + ATP sensory histidine kinase QseC - phosphorylated + ADP
UDP--D-glucuronate + glucosyl-(heptosyl)3-Kdo2-lipid A-phosphate UDP + glucosyl-(heptosyl)3-glucosyluronate-Kdo2-lipid A-phosphate
Qin prophage; cold shock protein
arnT
arnB
arnC
arnA
4-amino-4-deoxy-L-arabinose (L-Ara4N) transferase
undecaprenyl phosphate-L-Ara4FN transferase
UDP-L-Ara4O C-4" transaminase
-D-Kdo-(24)--D-Kdo-(26)-lipid A[periplasm] + 4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[periplasm] 4'--L-Ara4N--Kdo-(24)--Kdo-(26)-[P4'--L-ara4N]-lipid A[periplasm] + di-trans,octa-cis-undecaprenyl phosphate[periplasm]
fused UDP-L-Ara4N formyltransferase and UDP-GlcA dehydrogenase
UDP-4-amino-4-deoxy--L-arabinopyranose + 2-oxoglutarate UDP--L-threo-pentapyranos-4-ulose + L-glutamate
UDP-4-deoxy-4-formamido--L-arabinopyranose + di-trans,octa-cis-undecaprenyl phosphate 4-deoxy-4-formamido--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate + UDP
UDP-4-amino-4-deoxy--L-arabinopyranose + an N10-formyl-tetrahydrofolate UDP-4-deoxy-4-formamido--L-arabinopyranose + a tetrahydrofolate + H+
rpoC
UDP--D-glucuronate + NAD+ UDP--L-threo-pentapyranos-4-ulose + CO2 + NADH
waaG
RNA polymerase, subunit
polymyxin resistance
lipopolysaccharide glucosyltransferase I
Lipid A-core biosynthesis
a nucleoside triphosphate + RNA(n) RNA(n+1) + diphosphate
UDP--D-glucose + (heptosyl)2-Kdo2-lipid A glucosyl-(heptosyl)2-Kdo2-lipid A + UDP + H+
superpathway of lipopolysaccharide biosynthesis

Figure 3.1 *EVA-generated network.*

Network representation of mutated features that differ in parent strain, ML115, and evolved strains LAR1 and LAR2 and biological features related through gene regulation and metabolic pathways. Red nodes correspond to mutated features. Node shapes show feature type: ovals = genes, diamonds = gene products, triangles = reactions, hexagons = metabolic pathways.

RpoC, which contains a mutation both evolved strains, encodes the β' subunit of the RNA polymerase sigma 70 factor. The RNA polymerase sigma 70 factor is the primary sigma factor in *E. coli* K-12 MG1655 during exponential growth conditions (Jishage, Iwata, Ueda, & Ishihama, 1996) and functions to stabilize the open promoter complex during promoter melting and transcription initiation (Wigneshweraraj, Burrows, Severinov, & Buck, 2005). Thus, the mutation in *rpoC* gene could widely affect gene transcription in the evolved strains. Different mutations in *rpoC* gene were found in other evolutionary studies of acid tolerance, and the mutated RpoC (V507L) contributed to increased acid-tolerant phenotype (Harden et al., 2015). It is possible that the mutated RpoC (H419P) contributes to the increased C8 tolerance observed in the evolved strains through altered expression of genes with sigma 70 promoters.

We also confirmed a mutation in BasR (D28Y) in LAR1, which encodes the transcriptional regulator component of BasS-BasR system. The BasS-BasR two-component system is one of the two component signal transduction systems in *E. coli* which senses and responds to changes in environmental conditions (Hiroshi Ogasawara, Shota Shinohara, Kaneyoshi Yamamoto, & Akira Ishihama, 2012). In an evolutionary study of *n*-butanol tolerance, overexpression of *basS* was found to increase tolerance (Reyes, Almario, Winkler, Orozco, & Kao, 2012). Coincidentally, we found a 27 deletion in *basS* in LAR2, which encodes sensory histidine kinase of the BasS-BasR system.

The 768 bp insertion (InsB-5, InsA-5, and InsAB-5) found in *waaG* in the parent strain, ML115, is predicted to interrupt the expression of *waaG* and potentially alter the expression level of downstream genes in its operon: *waaP*, *waaS*, *waaS*, *waaO*, *waaJ*, *waaY*, *waaZ*, and *waaU* (Figure 3.2). In the evolved strains, this insertion was not detected, suggesting that the insertion sequence had moved and *waaG* had been restored. WaaG is a lipopolysaccharide (LPS) glucosyltransferase I enzyme which adds the first glucose of the outer core of LPS.

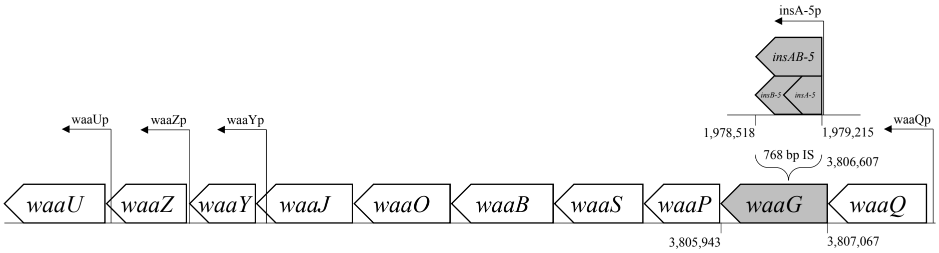


Figure 3.2 *An insertion sequence interrupts waaG in ML115.*

The insertion sequence, insAB-5, is found in the parent strain, ML115, and potentially affects transcription of other genes downstream of waaG.

Colony morphology of parent and evolved strains

The deletion of *waaG* gene has been shown to result in a truncated LPS core, loss of flagella pili (Parker et al., 1992), enhanced cell surface hydrophobicity, increased outer membrane permeability, and decreased ability of biofilm formation (Wang, Wang, Ren, Li, & Wang, 2015). The deletion of *waaGPBI* leads to a mucoid colony morphology (Parker et al., 1992), which is consistent with the morphological characteristics of ML115 compared to LAR1 on LB plates (Figure 3.3). Transmission electron microscopy images also reveal a lack of flagella in ML115 when *waaG* is non-functional and presence of flagella in LAR1 in which *waaG* is restored (Figure 3.4).

The order in which mutations were acquired during the metabolic evolution experiment

PCR experiments were performed on the parent strain, LAR1, and intermediate samples corresponding to the serial transfers performed in the original metabolic evolution experiment to detect the presence of the variant genomic sequence observed in the evolved strain. PCR fragments for each sample are shown in Figure 3.5. We find that *waaG* is repaired early in the experiment as can be seen by the decrease in fragment size which corresponds to the loss of the insertion sequence. The *rpoC* mutation is acquired in the middle of the metabolic evolution experiment, and the *basR* mutation is not detected until the end of the experiment. As LAR1 and LAR2 were not evolved independently, we may deduce that the *basS* mutation in LAR2 is similarly not acquired until the end of the experiment.

Growth ability of reconstructed strains in C8 challenge experiments

In order to identify mutations that contribute to C8 tolerance, we systematically introduced mutations into the parent strain, ML115, in the order they were acquired in the metabolic evolution experiment. We hypothesized that if a mutation was critical to C8 tolerance, its addition to ML115 would improve the growth rate in a C8 challenge experiment. In addition, the combined effect of the *basR* and *basS* mutations examined both with and without the repair of *waaG*.

None of the *rpoC*, *basS*, or *basR* mutations showed an improvement in C8 tolerance when introduced into ML115 individually that would account for the tolerance phenotype of LAR1 (Table 3.4). Removing the insertion sequence present in *waaG* in the parent strain increased the growth rate in 10 mM C8 compared to ML115 with the non-functional *waaG* gene (Figure 3.7A and B). The increase in growth ability observed in LAR1 vs ML115 (Figure 3.7A and E), however cannot be attributed to the restoration of *waaG* alone. The incorporation of the *rpoC* mutation after *waaG* is repaired further improves C8 tolerance (Figure 3.7C). Reconstructing the *basR* mutation after the *waaG* repair and *rpoC* mutation are incorporated does not further improve tolerance (Figure 3.7D).

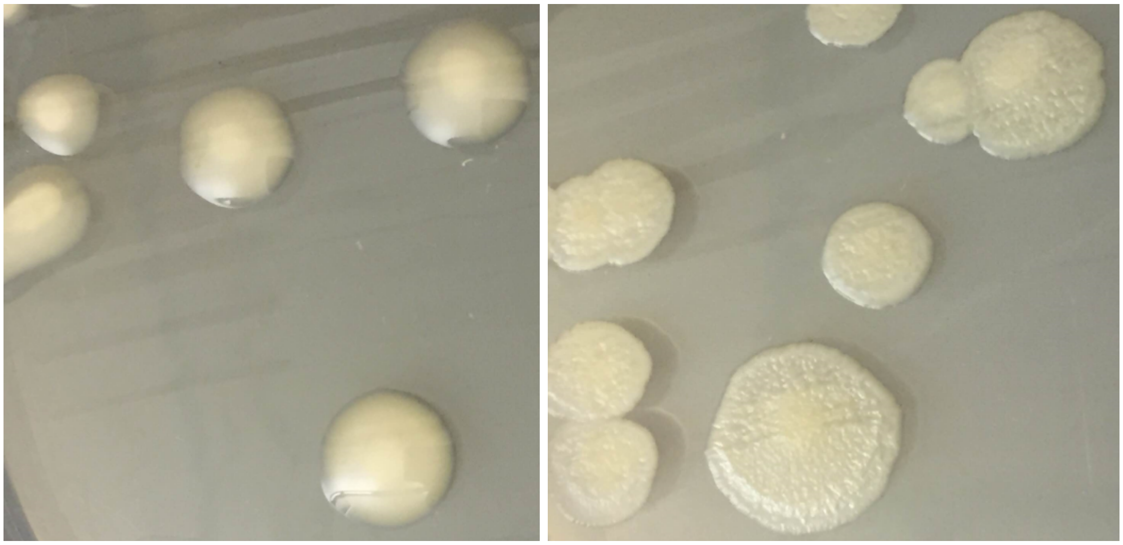


Figure 3.3 *Morphological characteristic of ML115 and LAR1.*

ML115 is shown on the left and LAR1 is shown on the right on LB agar plates at 37°C.

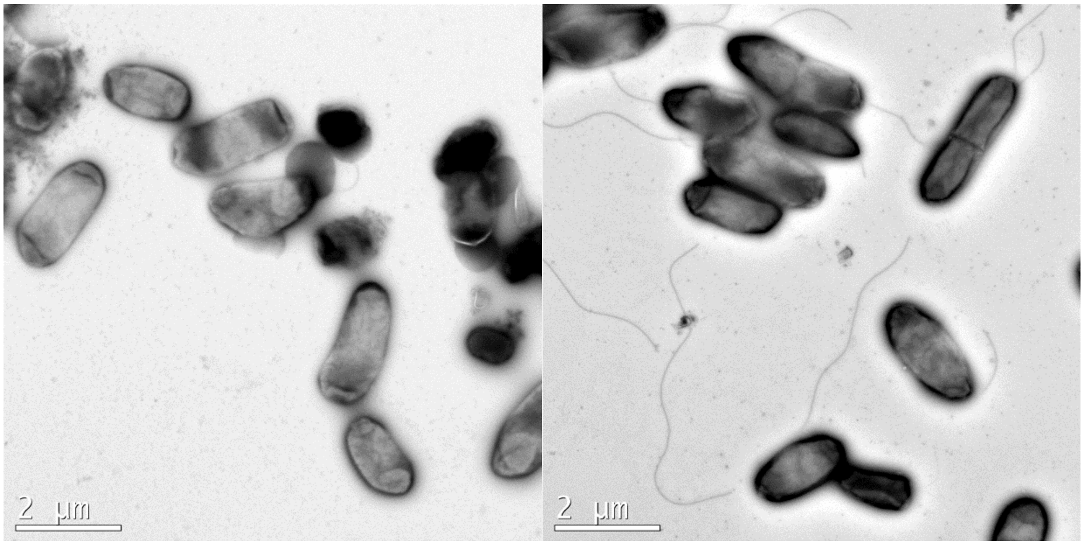


Figure 3.4 *Transmission electron microcopy images of ML115 and LAR1.*

ML115 is shown on the left and LAR1 is shown on the right. Flagella are noticeably absent in ML115 and restored in LAR1.

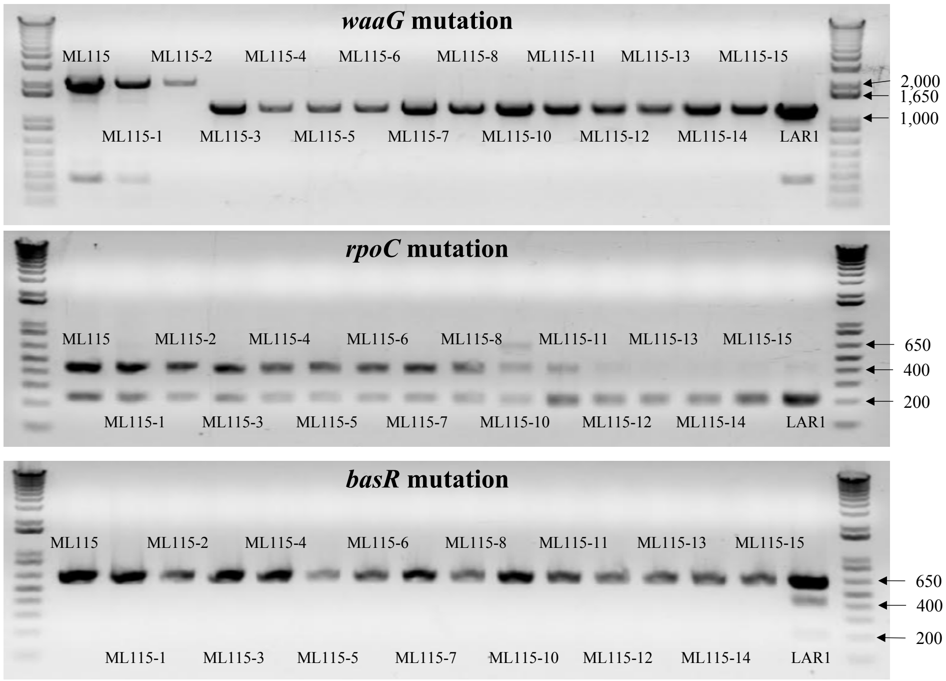


Figure 3.5 *PCR experiments to determine the order of mutations.*

PCR experiments for intermediate samples reveal the order in which mutations were acquired during the metabolic evolution experiment. Fragment sizes are labeled on the righthand ladder (in bp).

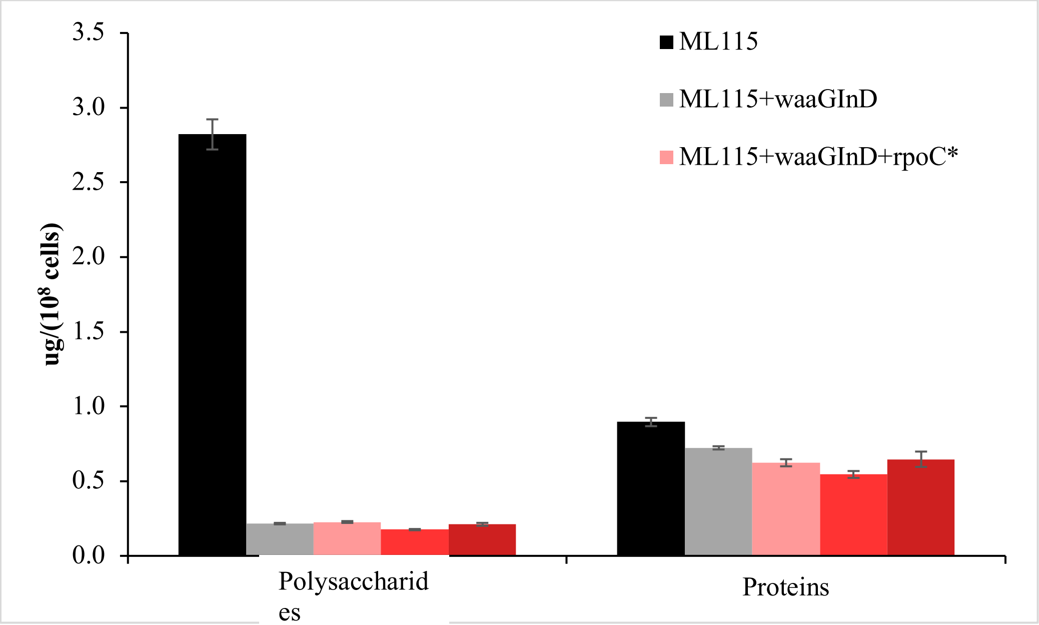


Figure 3.6 *Extracellular polymetric substance analysis for ML115, LAR1, and reconstructed strains.*

Interruption of *waaG* expression has been shown to result in a truncated LPS, decreased expression of major outer membrane proteins, and hypersensitivity to hydrophobic antibiotics (Parker et al., 1992; Yethon, Vinogradov, Perry, & Whitfield, 2000). In our previous study, the specific grow rate of MG1655 was greater than 0.5 h^{-1} in MOPS with 2.0% (w/v) glucose and 10 mM C8 (L. A. Royce et al., 2013), while the ML115 could barely grow under the same condition. Furthermore, octanoic acid is a hydrophobic chemical. Based on these observations, we believed the *waaG* insertion to cause octanoic acid hypersensitivity in the parent strain, ML115.

Interestingly, when measuring extracellular polymetric substances, we find an abundance of polysaccharides present in ML115 and a sharp decrease when *waaG* is restored (Figure 3.6). Many other genes are required to synthesize LPS and it is possible that the organism is compensating for the damaged *waaG* by overexpressing other LPS genes. The addition of other mutations does not greatly affect the abundance of polysaccharides. The protein content of the free extracellular polymetric substance was only slightly lower as the LAR1 genotype was reconstructed in the parent strain.

The repair of *rpoC* in LAR1 significantly decreased the growth rate in 10 mM C8 compared to the LAR1 strain (data not shown) and introducing the *rpoC* mutation into ML115 with a repaired *waaG* gene further increased C8 tolerance (Figure 3.7C). This indicates that the *rpoC* mutation is important for the C8 tolerance phenotype of evolved strain. As previously mentioned, the *rpoC* mutation could widely alter genes expression level.

Because of the rate of mutations in RNA polymerase genes in evolution experiments, there has been interest in investigating the relevance of these variations to fitness and their mechanisms (Conrad et al., 2010). For several mutations in *rpoB* and *rpoC* genes, Conrad et al. found improved growth in minimal media and slower growth in rich media. They also observed a decrease in the open complex longevity at the promoter and an increase in the transcript elongation rate. Transcriptomic analysis of ML115 and LAR1 is needed to uncover what global effects the rpoC mutation exerts on the evolved strain and which genes with altered expression might influence C8 tolerance.

The repair of *basR* in LAR1 strain and introduction in ML115 strain did not change the growth ability in the 10 mM C8 tolerance test (Figure 3.7D), demonstrating that the *basR* mutation alone is not able to increase the C8 tolerance. The introduction of the *basS* mutation from LAR2 into LAR1 did not further enhance the C8 tolerance phenotype. Similarly, the introduction of the *basS* mutation into ML115 did not affect tolerance at 10 mM C8 (Table 3.4). As the *basS* and *basR* mutations were acquired near the end of the evolution experiment, which corresponded to a concentration of 30 mM C8, repeating the growth experiment at a higher concentration may reveal some yet observed contribution to tolerance from these mutations.

The *waaG*, *basS* and *basR* mutations affect the cell membrane

Previous studies identified membrane damage as a key mechanism of microbial inhibition when applying exogenous octanoic acid challenge to *E. coli* strains or during fatty acid production (Jarboe et al., 2013; Lennen & Pfleger, 2013; L. A. Royce et al., 2013; L. A. Royce et al., 2015; Sherkhanov et al., 2014), but the introduction of *basS* and *basR* mutations in ML115 did not show improved tolerance at 10 mM C8. However, membrane characterization of the reconstructed strains revealed that these mutations improve cell membrane integrity.

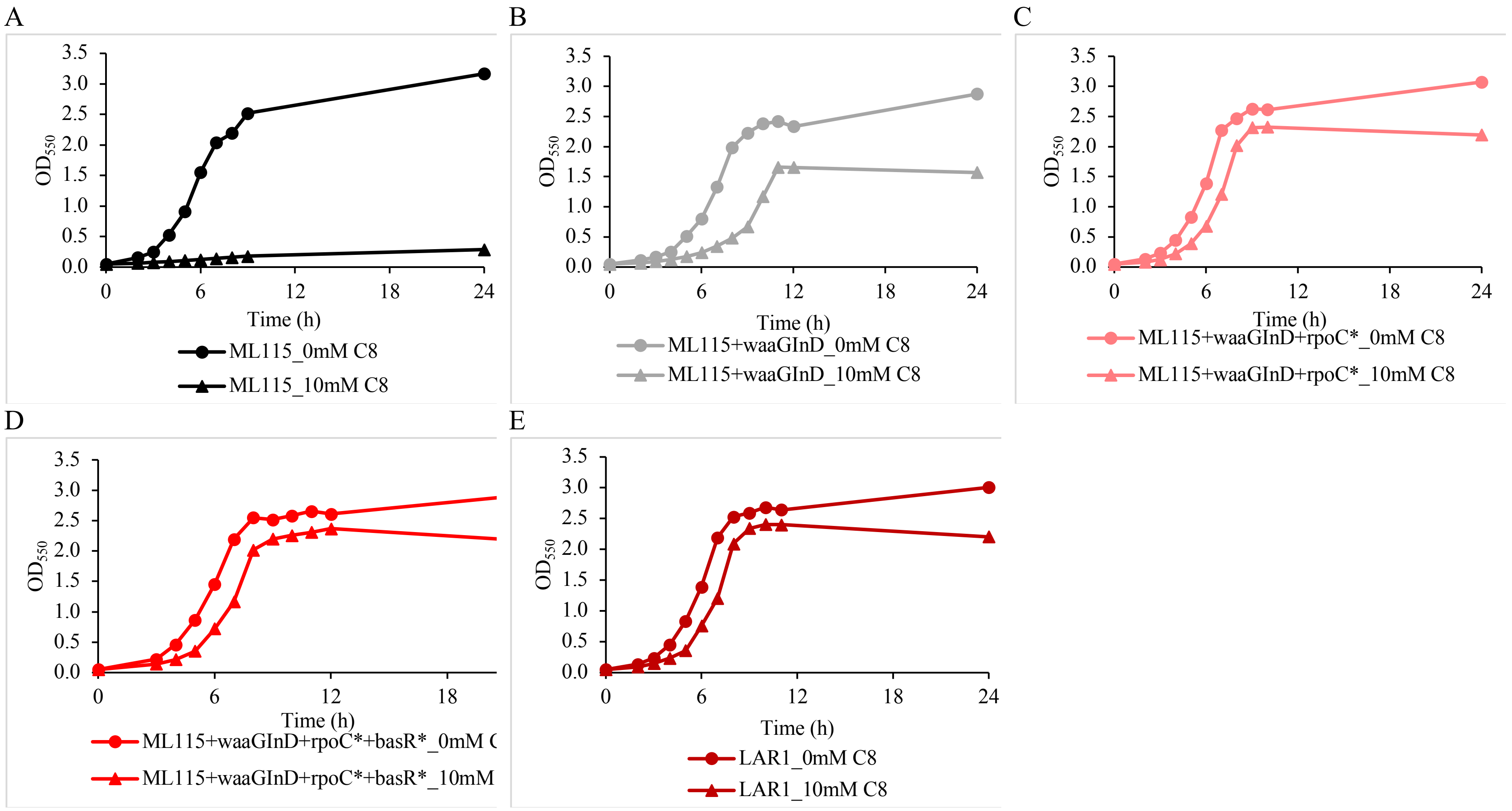


Figure 3.7 *Growth ability of ML115, LAR1, and reconstructed strains in the order that mutations were acquired.*

Strains were grown in MOPS media with 2.0% (w/v) glucose with 0 and 10 mM octanoic acid at 37°C. A. The parent strain, ML115; B. ML115 with waaG repaired; C. ML115 with repaired waaG and the rpoC mutation; D. ML115 with repaired waaG, the rpoC mutation, and the basR mutation; E. The evolved strain LAR1. Values are the average of three biological replicates. Coloring is consistent for these strains throughout all figures.

Table 3.4 *Growth ability of ML115, LAR1, and reconstructed strains.*

The top section of the table has data for the evolved strain, LAR1, the parent strain, ML115, and reconstructed strains with mutations in the order they were acquired in the metabolic evolution experiment. The bottom section of the table shows other reconstructed strains used to analyze independent and combined contributions of mutations to octanoic acid tolerance. For each strain, specific growth rate (GR), time at which log phase occurs, the inflection OD, and the 24 hour OD are provided. Coloring is consistent for the strains in all figures presented in this document. Inflection OD is defined as the value of OD_{550} and the first time point recorded corresponding to stationary phase.

Strain	**Gene**				**0 mM C8**				**10 mM C8**			
	waaG	*rpoC*	*basR*	*basS*	Specific GR	Log Phase	Inflection OD	24 h OD	Specific GR	Log Phase	Inflection OD	24 h OD
LAR1	*	*	*		0.56±0.00	3~7 h	2.53±0.07 (8 h)	3.01±0.04	0.56±0.00	4~8 h	2.34±0.04 (9 h)	2.20±0.07
ML115					0.60±0.01	3~6 h	2.04±0.07 (7 h)	3.17±0.07	0.15±0.00	ND	ND	0.29±0.01
ML115+waaGInD	*				0.53±0.01	3~7 h	1.99±0.02 (8 h)	2.87±0.17	0.39±0.01	6~11 h	1.66±0.03 (12 h)	1.57±0.03
ML115+waaGInD+rpoC*	*	*			0.57±0.01	3~7 h	2.47±0.00 (8)	3.07±0.11	0.55±0.02	4~8 h	2.31±0.06 (9 h)	2.20±0.02
ML115+waaGInD+rpoC*+basR*	*	*	*		0.57±0.00	3~7 h	2.55±0.03 (8 h)	2.98±0.04	0.57±0.01	4~8 h	2.20±0.03 (9 h)	2.13±0.01
ML115+rpoC*		*			0.65±0.01	3~6 h	1.95±0.04 (7 h)	3.00±0.02	0.23±0.01	ND	ND	0.38±0.02
ML115+basR*			*		0.60±0.00	3~6 h	1.90±0.02 (7 h)	2.61±0.02	0.15±0.00	ND	ND	0.28±0.01
ML115+basS*				*	0.60±0.01	3~6 h	1.96±0.04 (7 h)	2.74±0.02	0.16±0.00	ND	ND	0.31±0.01
ML115+waaGInD+basR*	*		*		0.45±0.01	4~8 h	2.39±0.02 (9 h)	3.00±0.10	0.35±0.01	6~11 h	1.92±0.02 (12 h)	1.57±0.07
ML115+waaGInD+basS*	*			*	0.52±0.01	3~7 h	2.25±0.03 (8 h)	2.95±0.12	0.35±0.01	5~9 h	1.56±0.08 (10 h)	1.30±0.08
ML115+rpoC*+basR*		*	*		0.67±0.00	2~5 h	1.98±0.01 (6 h)	2.88±0.05	0.16±0.00	ND	ND	0.17±0.02
ML115+basR*+basS*			*	*	0.54±0.00	3~7 h	1.94±0.04 (8 h)	2.34±0.09	0.20±0.01	ND	ND	0.21±0.01
ML115+waaGInD+basR*+basS*	*		*	*	0.47±0.01	3~8 h	2.44±0.03 (9 h)	2.98±0.04	0.39±0.00	4~10 h	1.76±0.02 (11 h)	1.53±0.03

Membrane fluidity can be measured as a fluorescence polarization, and the increased membrane polarization corresponds to decreased fluidity and an increase in membrane rigidity.

After being treated with 10 mM C8 for 1 hour, the ML115+basR* and ML115+basS* strains showed significantly higher membrane polarization than ML115 and ML115+rpoC* strains (Figure 3.8B), showing that *basS* and *basR* mutations improved cell membrane rigidity. There is no significant difference in cell membrane rigidity between LAR1 and LAR1+rpoC after treatment which showed that the *rpoC* mutation in LAR1 is not responsible for the increased cell membrane rigidity phenotype.

Directly comparing strains with and without a 10 mM C8 treatment revealed that all strains except ML115 and ML115+rpoC* have significantly higher membrane polarization, suggesting they can alter the cell membrane by sensing and responding to the altered environmental condition. Additionally, the ML115+basR*, ML115+basS* and LAR1+rpoC strains reached a similar membrane polarization level as LAR1 after treatment. We hypothesize that the *basS* and *basR* mutations are the key contribution to improve cell membrane rigidity in the evolved strains.

Examining reconstructed strains that illustrate the order of acquired mutations, we observe an increase in cell membrane rigidity at 10 mM C8 with the addition of each mutation (Figure 3.8A). There may exist some synergistic effect between the *waaG* repair and *rpoC* mutation. A further increase to cell membrane polarization is seen with the addition of the *basR* mutation at 10 mM C8.

Membrane leakage is another key factor of cell membrane damage. We used flow cytometry to separate and quantify SYTOX-permeable and SYOX-impermeable cells, where permeability of SYTOX indicates a damaged cell membrane. Without exogenous 10 mM C8 challenge, only about 5% of ML115, ML115+rpoC*, ML115+basR*, and ML115+basS* strain population became SYTOX-permeable (Figure 3.9A). After a C8 challenge for 1 hour, the permeable strain population of ML115 and ML115+rpoC* increased to 17.57±2.79% and 11.07±1.44%, while the permeable strain population of ML115+basR* and ML115+basS* decreased to 3.2±0.37% and 2.93±0.45% (Figure 3.9A).

These results show that restoring *waaG* is primarily responsible for preventing cell membrane leakage, the *rpoC* mutation partially restores cell membrane leakage, and the *basS* and *basR* mutations help reduce cells membrane leakage. Additionally, the SYTOX-permeable strain population of ML115+basR* and ML115+basS* significantly decreased after C8 challenge (Figure 3.9B), showing the altered cell membrane for strains with *basS* or *basR* mutations has increased the resistance to exogenous 10 mM C8 challenge compared to no C8 challenge. The SYTOX-permeable strain population of LAR1 and LAR1+rpoC were 2.17% and 2.7% without C8 treatment (Figure 3.9B). After treated with 10 mM C8 for 1 hour, the permeable strain population of LAR1 decreased to 1.93±0.45%, while it increased to 4.37±0.57% when the *rpoC* mutation repaired in LAR1 (Figure 3.9B).
This decreased resistance is consistent with the partially restores cell membrane leakage in ML115+rpoC*.

Strains that reconstruct the order of acquired mutations show that repairing *waaG* reduces the percent of SYTOX-permeable cells drastically. The next mutation, in *rpoC*, partially restores cell membrane leakage, and finally, the *basR* mutation again alleviates cell membrane leakage (Figure 3.9A).

Measurements of the membrane fluidity and membrane leakage help to understand how each mutation alter the cell membrane properties for increasing resistance to C8 tolerance and increasing fatty acid production. These results show that the *rpoC* mutation, while a factor of increasing C8 tolerance, does not contribute to increased membrane rigidity and partially restores cell membrane leakage. The *basS* and *basR* mutations can separately increase the cell membrane rigidity and prevent cell membrane leakage, but do not contribute to C8 tolerance phenotype at 10 mM C8.

Additionally, cell surface hydrophobicity was measured for individual mutations and strains that reconstruct the order of acquired mutations. The evolved strain, LAR1 exhibits a higher percentage of hydrophobicity than the parent strain, ML115 for which no individual mutation can account. The repair of *waaG* decreases the percent hydrophobicity and the *basS* mutation increases the percent hydrophobicity in 10 mM C8 compared to the control condition when each are introduced individually into ML115 (Figure 3.10B). However, when examining reconstructed strains with multiple mutations, an increase in percent hydrophobicity is seen with the addition of the *rpoC* and then *basR* mutations, suggesting a synergistic effect. The incorporation of the variant *basR* can achieve the percent hydrophobicity of the LAR1 strain.

Finally, we examine the effect of each mutation on membrane lipid composition as well as the cumulative effects of mutations in the order they were acquired. Interestingly, we find that different mutations result in different lipid compositions. For example, the *rpoC* mutation leads to increased mono-unsaturated fatty acids C16:1 and C18:1 and decreasd saturated fatty acid C16 while the restoration of *waaG* has the opposite effect (Figure 3.11A and B). The effects were similar with and without the 30 mM exogenous C8 treatment, but with an overall decrease in C14:0 and C17cyc. Examining mutations in the order they were acquired shows that mutations subsequent to *waaG* increased C18:1 and decreased C16:0 (Figure 3.11C and D).

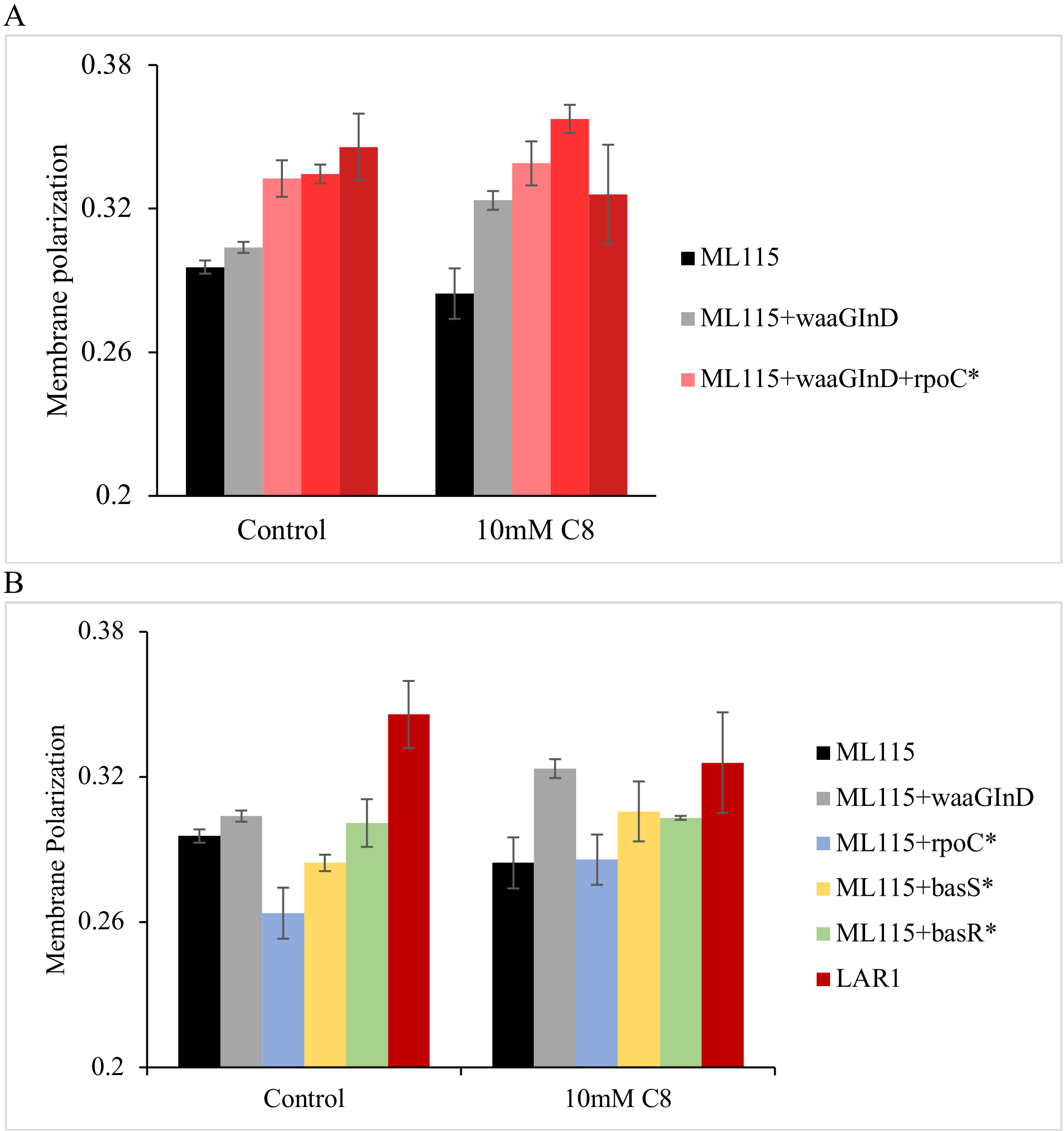


Figure 3.8 *Membrane polarization.*

Membrane polarization of reconstructed strains at mid-log phase (OD ≈ 1) with 10 mM C8 challenge. A. membrane polarization of strains in the order that mutations were acquired in the metabolic evolution experiment; B. membrane polarization for the parent strain, evolved strain, and strains with individual mutations introduced into ML115. Values are the average of three biological replicates, each biological replicate has four technical replicates and error bars show standard deviation.

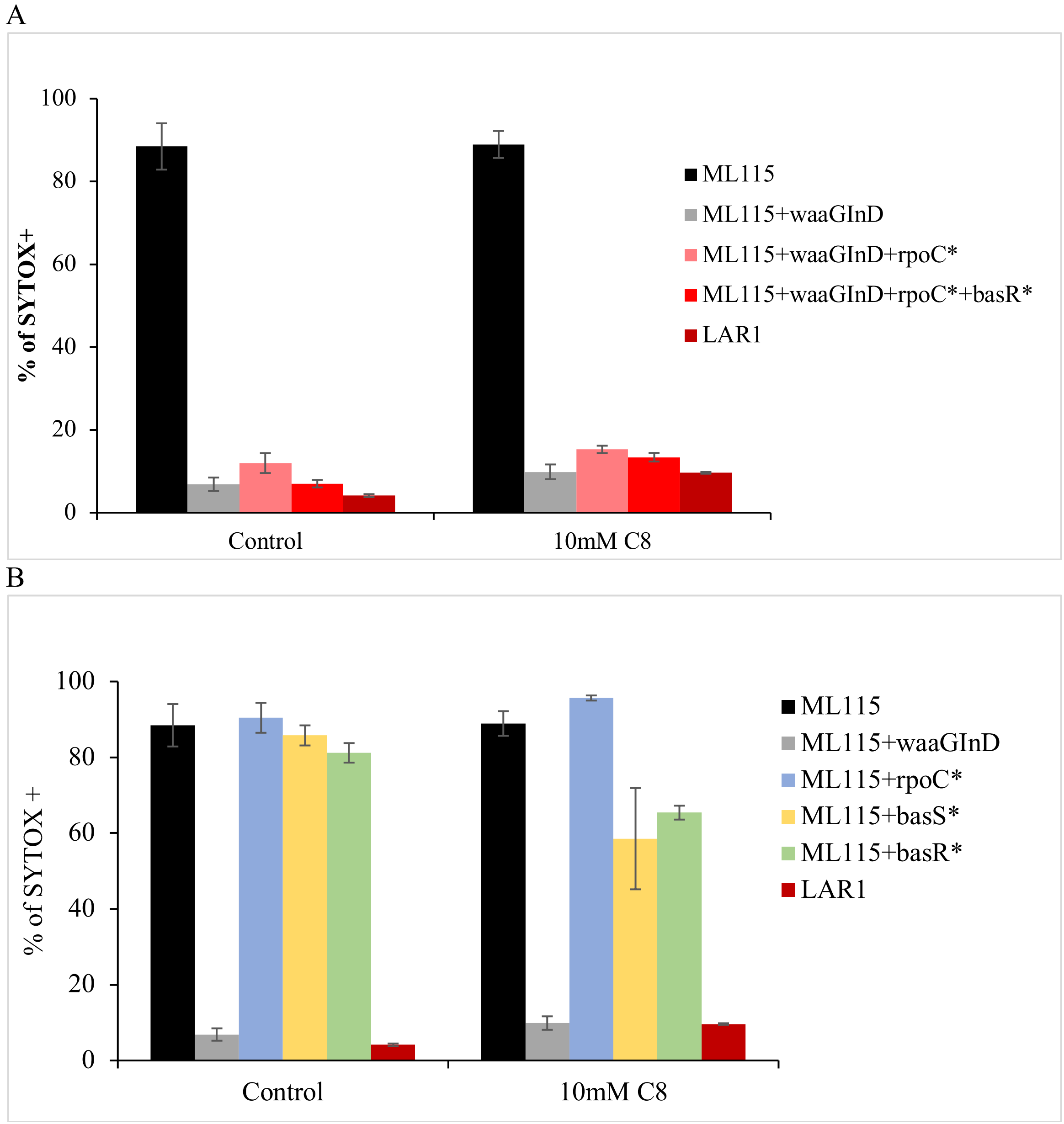


Figure 3.9 *Membrane leakage*

Membrane leakage of reconstructed strains and at mid-log phase (OD ≈ 1) with 10 mM C8 challenge. A. membrane leakage of strains that recreate the order in which mutations were acquired in the evolution experiment; B. membrane leakage of the parent strain, evolved strain LAR1, and individual mutations introduced into the parent strain. Values are the average of three biological replicates, and error bars show standard deviation.

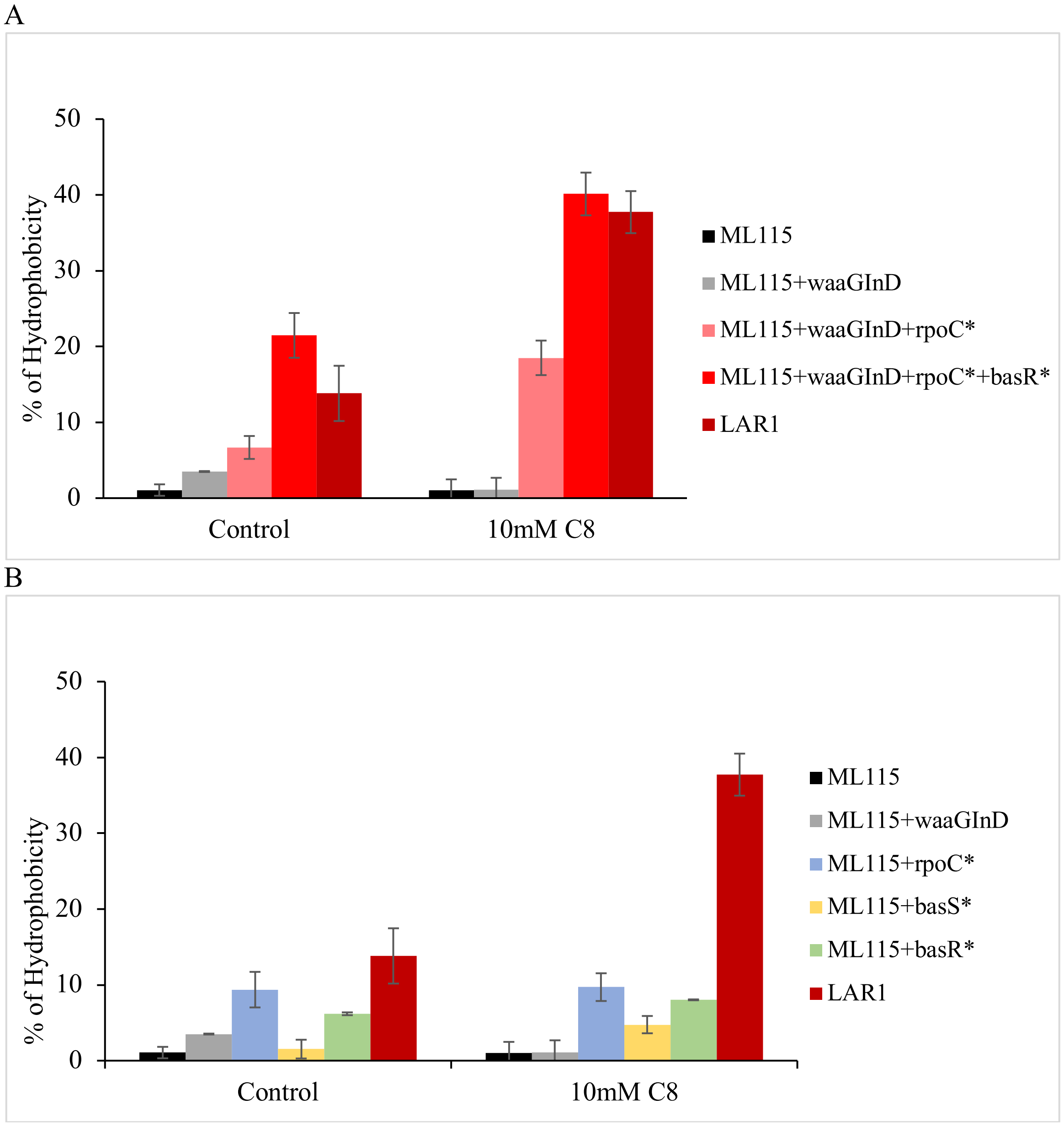


Figure 3.10 *Percent of hydrophobicity.*

Percent of hydrophobicity for parent strain, evolved strain, and other reconstructed strains. A. reconstructed strains that recreate the order in which mutations were acquired in the metabolic evolution experiment; B. reconstructed strains illustrating individual contribution of mutations. Values are the average of two biological replicates and two technical replicates and error bars show standard deviation.

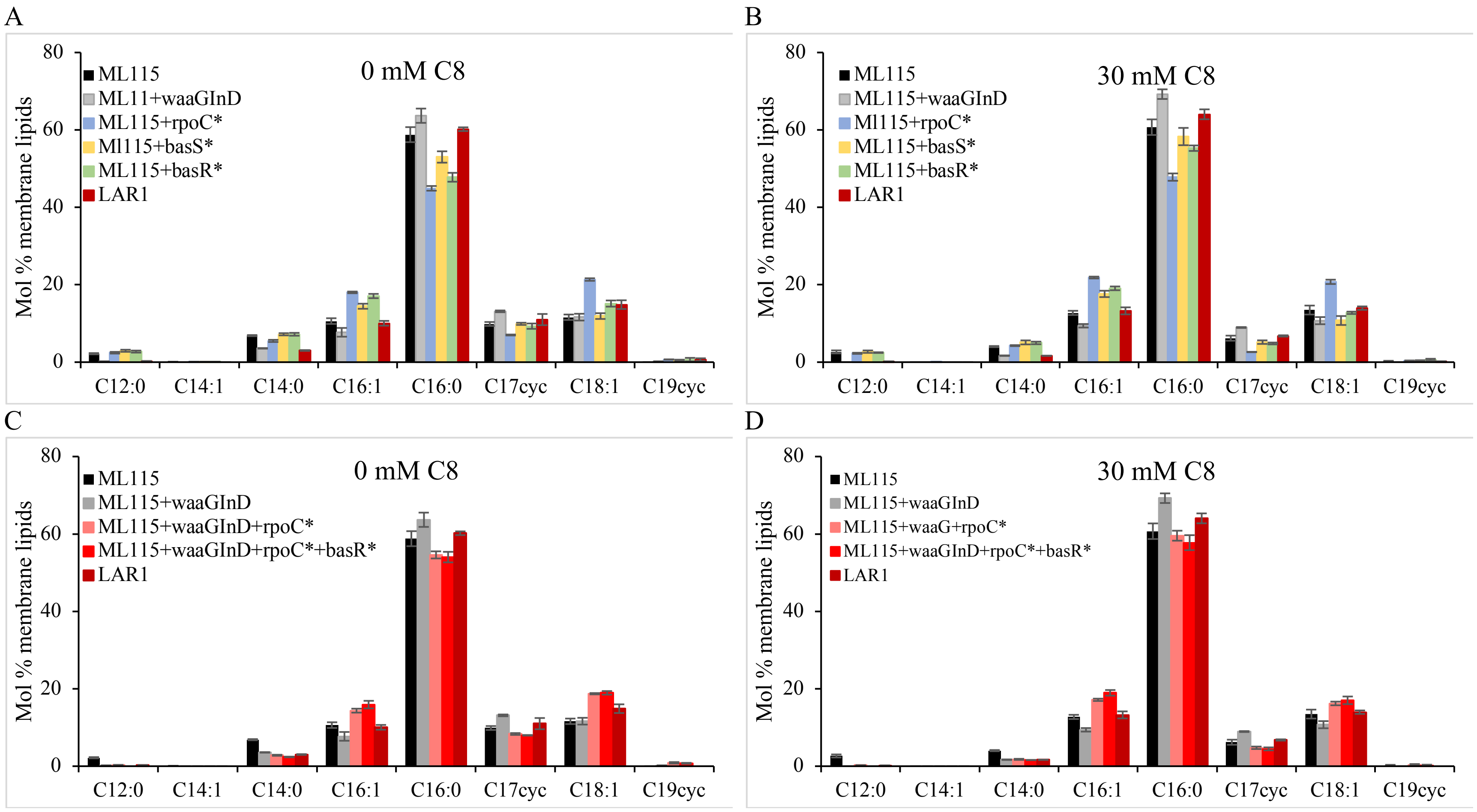


Figure 3.11 *Membrane lipid composition.*

Membrane lipid composition for reconstructed strains at 0 (A and C) and 30 mM (B and D) C8. Values are the averages of three biological replicates and error bars show standard deviation. A and B show the effects of individual mutations on membrane lipid composition and C and D show the cumulative effects of mutations acquired in the metabolic evolution experiment.

The *waaG* and *rpoC* mutations improve fatty acid titer

In addition to improving tolerance to octanoic acid, the repair of *waaG* demonstrates some improvement to fatty acid titer and when the introduction of the *rpoC* mutation is incorporated, titer matching that of LAR1 at 24 and 72 hours is observed (Figure 3.12). We have shown that a non-functional *waaG* decreases C8 tolerance and alters membrane properties. The mutation in *rpoC*, the β' subunit of RNA polymerase sigma 70 subunit, is anticipated to alter expression of genes with sigma 70-associated promoters in exponential growth conditions. We previously described the mutation as predicted to be damaging in EVA analysis because the genomic variation is not found among published genomic sequences, however, as *rpoC* plays a critical role in transcription initiation, it is likely that its behavior is modified, which in turn, could achieve a wide range of phenotypic effects by perturbing global transcription. Indeed, mutations in the primary RNA polymerase genes have been found in several adaptive evolution experiments (Applebee, Herrgård, & Palsson, 2008; Jin & Gross, 1988; Klein-Marcuschamer, Santos, Yu, & Stephanopoulos, 2009; Trinh, Langelier, Archambault, & Coulombe, 2006; Zhou & Jin, 1998). In addition to increasing the specific growth rate in the C8 challenge condition, the *rpoC* mutation appears to be a critical mutation for rewiring global transcription to support increased fatty acid titer in LAR1.

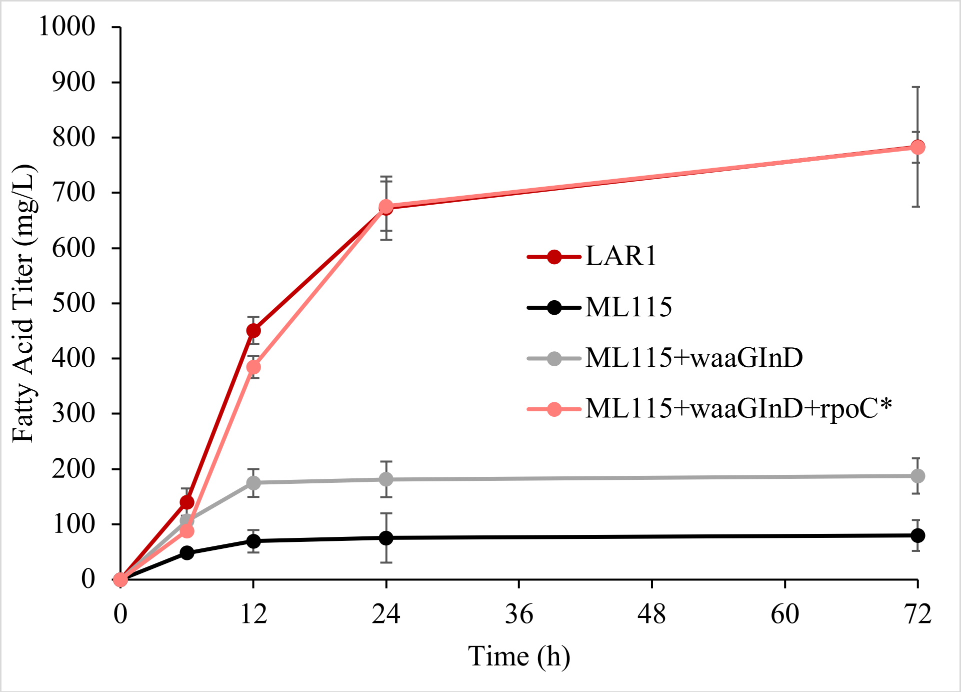


Figure 3.12 *Fatty acid titer measured over time.*

Fatty acid titers for ML115, LAR1, and reconstructed strains at 6, 12, 24, and 72 hours. Values are the average of three biological replicates and error bars represent standard deviation.

Summary

In this study we reverse engineered *E. coli* strains LAR1 and LAR2 which were evolved for improved octanoic acid tolerance. Through whole genome sequencing, we identified mutations in *rpoC* in both evolved strains and *basR* and *basS* separately in each strain. We also discovered a key mutation in the parent strain, ML115, which was an insertion sequence in the *waaG* gene that was lost during the experiment. Through PCR analysis of intermediate samples, we showed that waaG was repaired early in the evolution experiment, followed by the *rpoC* mutation, and finally the *basS* and *basR* mutations in LAR1 and LAR2, respectively. We predict each of these mutations to be damaging based on

Provean analysis which considers conservation at the mutation site among published gene products with sequence similarity.

Both individual and collections of mutations were repaired in the evolved strain and introduced in the parent strain in order to characterize their contribution to phenotype. We show that non-functional *waaG* is detrimental to tolerance and affects membrane properties. A drastic reduction in membrane leakage is observed after *waaG* is repaired. The *basS* and *basR* mutations alter the cell membrane by increasing percent hydrophobicity and membrane polarization. Finally, the *rpoC* mutation contributes both to tolerance and increases fatty acid titer.

The independent mutations in *basS* and *basR* may affect the BasS-BasR two-component signal transduction system and alter transcription of genes in the BasR regulon. Using the gene regulatory and metabolic network generated by EVA, we identified candidate genes in the BasR regulon *waaH* and *eptA* for further study based on their function in modifying LPS. The *csg* operon was also of interest for its role in curli assembly and biofilm formation. Modification of these functions may have a similar effect on the evolved strain phenotype as the repair of waaG which is also involved in LPS modification, surface organelle biosynthesis, and biofilm formation.

We hypothesize that the *rpoC* mutation alters global transcription in the evolved strains. We searched promoter sequences in the publicly available database, RegulonDB, and found 1,606 genes with associated sigma 70 promoters. Due to this large number of genes and sigma 70 being the primary sigma factor during exponential growth, a forthcoming RNA-seq study will compare the transcriptomes of ML115 and LAR1 and identify genes with altered expression that contribute to the tolerance phenotype.

Acknowledgements

The authors would like to thank Michael Baker and the rest of the Iowa State University DNA Facility.

Authors' Contributions

EEB performed DNA sequence alignment, mutation prediction, and mutation analysis. YC constructed the strains described in this study, performed growth and titer experiments, and further characterized strains through membrane and EPS studies. The project was conceived and advised by JAD and LRJ. All authors read and reviewed the final version of the manuscript.

References

Applebee, M. K., Herrgård, M. J., & Palsson, B. (2008). Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. J Bacteriol, 190(14), 5087-5094. doi:10.1128/JB.01976-07

Atsumi, S., Wu, T.-Y., Machado, I. M. P., Huang, W.-C., Chen, P.-Y., Pellegrini, M., & Liao, J. C. (2010). Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Molecular Systems Biology, 6(1), 449. doi:papers3://publication/uuid/E5E4CC66-CC2F-4343-8E7C-32F52AA61D28

Baer, S. H., Blaschek, H. P., & Smith, T. L. (1987). Effect of Butanol Challenge and Temperature on Lipid Composition and Membrane Fluidity of Butanol-Tolerant Clostridium acetobutylicum. Applied and Environmental Microbiology, 53(12), 2854-2861. doi:papers3://publication/uuid/424000F3-CEC9-4CBB-9016-7E4968DDE675

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., . . . Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. Science, 277(5331), 1453-1462. doi:papers3://publication/uuid/5530EBDB-8A2B-4B0F-8B27-8116B7DB359F

Bligh, E. G., & Dyer, W. J. (1959). A rapid method of total lipid extraction and purification. Can J Biochem Physiol, 37(8), 911-917. doi:10.1139/o59-099

Boggess, E., Jarboe, L., & Dickerson, J. (2018). Mutation analysis for metabolic evolution experiments in Escherichia coli. BMC Bioinformatics, Submitted.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino Acid substitutions and Indels. PLoS ONE, 7(10), e46688. doi:papers3://publication/uuid/53E1A263-4B6F-4386-A543-A5323B4F51BF

Conrad, T. M., Frazier, M., Joyce, A. R., Cho, B. K., Knight, E. M., Lewis, N. E., . . . Palsson, B. (2010). RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. Proc Natl Acad Sci U S A, 107(47), 20500-20505. doi:10.1073/pnas.0911253107

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6640-6645. doi:papers3://publication/doi/10.1073/pnas.120163297

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods in molecular biology (Clifton, N.J.), 1151, 165-188. doi:papers3://publication/uuid/68DB31A8-324E-4B2A-BE23-7988DC930923

Dellomonaco, C., Fava, F., & Gonzalez, R. (2010). The path to next generation biofuels: successes and challenges in the era of synthetic biology. Microbial cell factories, 9(1), 3. doi:papers3://publication/doi/10.1186/1475-2859-9-3

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., . . . Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human mutation, 37(6), 564-569. doi:papers3://publication/uuid/C7D577F4-78B5-48BA-BDF4-741B9EA1FAC6

Froelich, J. M., Tran, K., & Wall, D. (2006). A pmrA constitutive mutant sensitizes Escherichia coli to deoxycholic acid. J Bacteriol, 188(3), 1180-1183. doi:10.1128/JB.188.3.1180-1183.2006

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic acids research, 44(D1), D133-143. doi:papers3://publication/uuid/59309EE8-734B-4AAF-943A-0C4E0BA72C35

Harden, M. M., He, A., Creamer, K., Clark, M. W., Hamdallah, I., Martinez, K. A., . . . Slonczewski, J. L. (2015). Acid-adapted strains of Escherichia coli K-12 obtained by experimental evolution. Applied and Environmental Microbiology, 81(6), 1932-1941. doi:papers3://publication/uuid/FD16E8AE-1E58-4F5A-BB32-38557CEBDEA3

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., . . . Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol Syst Biol, 2, 2006.0007. doi:10.1038/msb4100049

Huffer, S., Roche, C. M., Blanch, H. W., & Clark, D. S. (2012). Escherichia coli for biofuel production: bridging the gap from promise to practice. Trends in biotechnology, 30(10), 538-545. doi:papers3://publication/doi/10.1016/j.tibtech.2012.07.002

Jarboe, L. R., Liu, P., & Royce, L. A. (2011). Engineering inhibitor tolerance for the production of biorenewable fuels and chemicals. Current Opinion in Chemical Engineering, 1(1), 38-42. doi:papers3://publication/doi/10.1016/j.coche.2011.08.003

Jarboe, L. R., Royce, L. A., & Liu, P. (2013). Understanding biocatalyst inhibition by carboxylic acids. Frontiers in microbiology, 4, 272. doi:papers3://publication/doi/10.3389/fmicb.2013.00272

Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J., & Yang, S. (2015). Multigene editing in the Escherichia coli genome via the CRISPR-Cas9 system. Applied and Environmental Microbiology, 81(7), 2506-2514. doi:papers3://publication/doi/10.1128/AEM.04023-14

Jin, D. J., & Gross, C. A. (1988). Mapping and sequencing of mutations in the Escherichia coli rpoB gene that lead to rifampicin resistance. J Mol Biol, 202(1), 45-58.

Jishage, M., Iwata, A., Ueda, S., & Ishihama, A. (1996). Regulation of RNA polymerase sigma subunit synthesis in Escherichia coli: intracellular levels of four species of sigma subunit under various growth conditions. J Bacteriol, 178(18), 5447-5451.

Klein-Marcuschamer, D., Santos, C. N., Yu, H., & Stephanopoulos, G. (2009). Mutagenesis of the bacterial RNA polymerase alpha subunit for improvement of complex phenotypes. Appl Environ Microbiol, 75(9), 2705-2711. doi:10.1128/AEM.01888-08

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359. doi:papers3://publication/doi/10.1038/nmeth.1923

Lennen, R. M., Kruziki, M. A., Kumar, K., Zinkel, R. A., Burnum, K. E., Lipton, M. S., . . . Pfleger, B. F. (2011). Membrane stresses induced by overproduction of free fatty acids in Escherichia coli. Applied and Environmental Microbiology, 77(22), 8114-8128. doi:papers3://publication/doi/10.1128/AEM.05421-11

Lennen, R. M., & Pfleger, B. F. (2012). Engineering Escherichia coli to synthesize free fatty acids. Trends in biotechnology, 30(12), 659-667. doi:papers3://publication/doi/10.1016/j.tibtech.2012.09.006

Lennen, R. M., & Pfleger, B. F. (2013). Modulating membrane composition alters free fatty acid tolerance in Escherichia coli. PLoS ONE, 8(1), e54031. doi:papers3://publication/doi/10.1371/journal.pone.0054031

Miller, E. N., Jarboe, L. R., Turner, P. C., Pharkya, P., Yomano, L. P., York, S. W., . . . Ingram, L. O. (2009). Furfural inhibits growth by limiting sulfur assimilation in ethanologenic Escherichia coli strain LY180. Applied and Environmental Microbiology, 75(19), 6132-6141. doi:papers3://publication/doi/10.1128/AEM.01187-09

Miller, E. N., Jarboe, L. R., Yomano, L. P., York, S. W., Shanmugam, K. T., & Ingram, L. O. (2009). Silencing of NADPH-dependent oxidoreductase genes (yqhD and dkgA) in furfural-resistant ethanologenic Escherichia coli. Applied and Environmental Microbiology, 75(13), 4315-4323. doi:papers3://publication/doi/10.1128/AEM.00567-09

Mykytczuk, N. C. S., Trevors, J. T., Leduc, L. G., & Ferroni, G. D. (2007). Fluorescence polarization in studies of bacterial cytoplasmic membrane fluidity under environmental stress. Progress in biophysics and molecular biology, 95(1-3), 60-82. doi:papers3://publication/doi/10.1016/j.pbiomolbio.2007.05.001

Nikolau, B. J., Perera, M. A. D. N., Brachova, L., & Shanks, B. (2008). Platform biochemicals for a biorenewable chemical industry. The Plant journal : for cell and molecular biology, 54(4), 536-545. doi:papers3://publication/doi/10.1111/j.1365-313X.2008.03484.x

Ogasawara, H., Shinohara, S., Yamamoto, K., & Ishihama, A. (2012). Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. Microbiology (Reading, England), 158(Pt 6), 1482-1492. doi:papers3://publication/doi/10.1099/mic.0.057745-0

Ogasawara, H., Shinohara, S., Yamamoto, K., & Ishihama, A. (2012). Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. Microbiology, 158(Pt 6), 1482-1492. doi:10.1099/mic.0.057745-0

Parker, C. T., Kloser, A. W., Schnaitman, C. A., Stein, M. A., Gottesman, S., & Gibson, B. W. (1992). Role of the rfaG and rfaP genes in determining the lipopolysaccharide core structure and cell surface properties of Escherichia coli K-12. J Bacteriol, 174(8), 2525-2538.

Pembrey, R. S., Marshall, K. C., & Schneider, R. P. (1999). Cell surface analysis techniques: What do cell preparation protocols do to cell surface properties? Applied and Environmental Microbiology, 65(7), 2877-2894. doi:papers3://publication/uuid/BD9E3A50-AE8F-4ACF-970B-D63406905D36

Perez, J. M., Richter, H., Loftus, S. E., & Angenent, L. T. (2013). Biocatalytic reduction of short-chain carboxylic acids into their corresponding alcohols with syngas fermentation. Biotechnology and bioengineering, 110(4), 1066-1077. doi:papers3://publication/doi/10.1002/bit.24786

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.4.1). Vienna, Austria: R Foundation for Statistical Computing

Reyes, L. H., Almario, M. P., Winkler, J., Orozco, M. M., & Kao, K. C. (2012). Visualizing evolution in real time to determine the molecular mechanisms of n-butanol tolerance in Escherichia coli. Metabolic Engineering, 14(5), 579-590. doi:papers3://publication/doi/10.1016/j.ymben.2012.05.002

Royce, L., Boggess, E., Jin, T., Dickerson, J., & Jarboe, L. (2013). Identification of mutations in evolved bacterial genomes. Methods in molecular biology (Clifton, N.J.), 985, 249-267. doi:papers3://publication/doi/10.1007/978-1-62703-299-5_13

Royce, L. A., Boggess, E., Fu, Y., Liu, P., Shanks, J. V., Dickerson, J., & Jarboe, L. R. (2014). Transcriptomic analysis of carboxylic acid challenge in Escherichia coli: beyond membrane damage. PLoS ONE, 9(2), e89580. doi:papers3://publication/doi/10.1371/journal.pone.0089580

Royce, L. A., Liu, P., Stebbins, M. J., Hanson, B. C., & Jarboe, L. R. (2013). The damaging effects of short chain fatty acids on Escherichia coli membranes. Applied microbiology and biotechnology, 97(18), 8317-8327. doi:papers3://publication/doi/10.1007/s00253-013-5113-5

Royce, L. A., Yoon, J. M., Chen, Y., Rickenbach, E., Shanks, J. V., & Jarboe, L. R. (2015). Evolution for exogenous octanoic acid tolerance improves carboxylic acid production and membrane integrity. Metabolic Engineering, 29, 180-188. doi:papers3://publication/doi/10.1016/j.ymben.2015.03.014

Sherkhanov, S., Korman, T. P., & Bowie, J. U. (2014). Improving the tolerance of Escherichia coli to medium-chain fatty acid production. Metabolic Engineering, 25, 1-7. doi:papers3://publication/doi/10.1016/j.ymben.2014.06.003

Stephanopoulos, G. (2007). Challenges in engineering microbes for biofuels production. Science, 315(5813), 801-804. doi:papers3://publication/doi/10.1126/science.1139612

Torella, J. P., Ford, T. J., Kim, S. N., Chen, A. M., Way, J. C., & Silver, P. A. (2013). Tailored fatty acid synthesis via dynamic control of fatty acid elongation. Proceedings of the National Academy of Sciences, 110(28), 11290-11295. doi:papers3://publication/doi/10.1073/pnas.1307129110

Trinh, V., Langelier, M. F., Archambault, J., & Coulombe, B. (2006). Structural perspective on mutations affecting the function of multisubunit RNA polymerases. Microbiol Mol Biol Rev, 70(1), 12-36. doi:10.1128/MMBR.70.1.12-36.2006

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Res, 40(15), e115. doi:10.1093/nar/gks596

Volker, A. R., Gogerty, D. S., Bartholomay, C., Hennen-Bierwagen, T., Zhu, H., & Bobik, T. A. (2014). Fermentative production of short-chain fatty acids in Escherichia coli. Microbiology (Reading, England), 160(Pt 7), 1513-1522. doi:papers3://publication/doi/10.1099/mic.0.078329-0

Wang, Z., Wang, J., Ren, G., Li, Y., & Wang, X. (2015). Influence of Core Oligosaccharide of Lipopolysaccharide to Outer Membrane Behavior of Escherichia coli. Mar Drugs, 13(6), 3325-3339. doi:10.3390/md13063325

Wigneshweraraj, S. R., Burrows, P. C., Severinov, K., & Buck, M. (2005). Stable DNA opening within open promoter complexes is mediated by the RNA polymerase beta'-jaw domain. J Biol Chem, 280(43), 36176-36184. doi:10.1074/jbc.M506416200

Wilmes-Riesenberg, M. R., & Wanner, B. L. (1992). TnphoA and TnphoA' elements for making and switching fusions for study of transcription, translation, and cell surface localization. Journal of bacteriology, 174(14), 4558-4575. doi:papers3://publication/doi/10.1128/jb.174.14.4558-4575.1992

Wu, H., Karanjikar, M., & San, K.-Y. (2014). Metabolic engineering of Escherichia coli for efficient free fatty acid production from glycerol. Metabolic Engineering, 25, 82-91. doi:papers3://publication/doi/10.1016/j.ymben.2014.06.009

Wu, H., Lee, J., Karanjikar, M., & San, K.-Y. (2014). Efficient free fatty acid production from woody biomass hydrolysate using metabolically engineered Escherichia coli. Bioresource technology, 169, 119-125. doi:papers3://publication/doi/10.1016/j.biortech.2014.06.092

Yethon, J. A., Vinogradov, E., Perry, M. B., & Whitfield, C. (2000). Mutation of the lipopolysaccharide core glycosyltransferase encoded by waaG destabilizes the outer membrane of Escherichia coli by interfering with core phosphorylation. J Bacteriol, 182(19), 5620-5623.

Yomano, L. P., York, S. W., & Ingram, L. O. (1998). Isolation and characterization of ethanol-tolerant mutants of Escherichia coli KO11 for fuel ethanol production. Journal of industrial microbiology & biotechnology, 20(2), 132-138. doi:papers3://publication/uuid/3B07DF98-8E97-4AA8-9F62-10348629BF0A

Zhang, C., Yang, H., Yang, F., & Ma, Y. (2009). Current Progress on Butyric Acid Production by Fermentation. Current Microbiology, 59(6), 656-663. doi:papers3://publication/doi/10.1007/s00284-009-9491-y

Zhang, X., Jantama, K., Moore, J. C., Jarboe, L. R., Shanmugam, K. T., & Ingram, L. O. (2009). Metabolic evolution of energy-conserving pathways for succinate production in Escherichia coli. Proceedings of the National Academy of Sciences, 106(48), 20180-20185. doi:papers3://publication/doi/10.1073/pnas.0905396106

Zhou, Y. N., & Jin, D. J. (1998). The rpoB mutants destabilizing initiation complexes at stringently controlled promoters behave like "stringent" RNA polymerases in Escherichia coli. Proc Natl Acad Sci U S A, 95(6), 2908-2913.

CHAPTER 4. TRANSCRIPTOMIC ANALYSIS OF *ESCHERICHIA COLI* EVOLVED FOR OCTANOIC ACID TOLERANCE

A research paper to be submitted to Metabolic Engineering

Erin E. Boggess[1], Yingxi Chen[2], Laura R. Jarboe[2], Julie A. Dickerson[1]

[1]*Department of Electrical & Computer Engineering, Iowa State University, Ames, IA, 50011.*

[2]*Department of Chemical & Biological Engineering, Iowa State University, Ames, IA, 50011.*

Abstract

Reverse engineering of strains obtained through metabolic evolution remains a significant challenge. Whole genome analysis is crucial for identifying mutations that are responsible for altered phenotype. These mutations can be introduced into ancestral strains and repaired in evolved strains to determine their relevance to improved fitness. In some cases, however, mutations arise in regulators and to understand the biological mechanisms responsible for phenotype, regulated genes must be studied. In our previous work, we identify mutations in a transcription factor and the β' subunit of RNA polymerase (*rpoC*) in *Escherichia coli* evolved for improved octanoic acid tolerance (Boggess, Jarboe, & Dickerson, 2018). Here, we present an RNA-seq study to support reverse engineering efforts and integrate our transcriptomic analysis findings with gene regulatory and metabolic pathway data. We identify differentially expressed genes regulated by mutated features as candidate genes, construct gene knockout strains, and test for altered growth rate with exogenous octanoic acid.

Introduction

Metabolic evolution been previously used as a successful strategy for developing strains of *Escherichia coli* with increased improved to octanoic acid and increased short-chain fatty acid production (C8) (Royce et al., 2015). These strains have been characterized has having an altered membrane with increased polarization, decreased leakage, longer average lipid length, and an altered saturated to unsaturated fatty acid ratio. The evolved strains also have restored flagella, a decrease in extracellular polymetric substances and an increase in percentage hydrophobicity (Chen, Boggess, Dickerson, & Jarboe, 2018). Our previous work in reverse engineering the evolved strains LAR1 and LAR2 involved whole genome sequencing and mutation analysis. We identified an insertion sequence that leads to a non-functional *waaG* in the parent strain, ML115, a point mutation in *rpoC* in both evolved strains, a point mutation in *basR* in LAR1, and a 27 base pair (bp) deletion in *basS* in LAR2 (Chen et al., 2018).

Our previous work identified the restoration of *waaG* as a large contributor to the evolved strain phenotype, however this mutation does not reproduce the fatty acid titer and growth ability observed in LAR1. The mutations involving RpoC, the β’ subunit of RNA polymerase, and BasS-R, a two-component signal transduction system are believed to affect expression of other genes through transcription regulation rather than directly contribute to phenotype. The variant β’ subunit may affect global gene expression in the evolved strains. The BasR regulon involves 22 genes, however some of these genes encode transcription factors and expression for additional downstream genes in the regulatory network may be affected by the mutant BasR.

In order to continue our work on reverse engineering LAR1, we must investigate the effects of mutations in regulators. In this work, we demonstrate the added value of omics experiments for reverse engineering microbial strains produced in metabolic evolution experiments, particularly when variant global regulators and transcription factors are found in evolved strains. We also integrate transcriptomic data into the gene regulatory and metabolic network we hypothesize are affected by genomic mutations (Boggess et al., 2018). We performed RNA-seq experiments for ML115 and LAR1 for control and fatty acid production conditions. We analyzed transcriptomic data to find genes that are differentially expressed in the parent and evolved strain. Incorporating our prior knowledge about the genotypes of these strains and integrating transcriptomic data with gene regulatory and metabolic pathway data, we identified candidate genes to test for relevance to improved phenotype.

Methods

Bacterial strains, plasmids, and media

Plasmids, strains from the octanoic acid evolution experiment, and reconstructed strains from this study are described in **Error! Reference source not found.**. *E. coli* DH5α was used as a cloning strain, while the parent strain, ML115 and the evolved strain LAR1 were used in genome modification procedures. All strains were grown overnight at 37ºC with 250 rpm orbital shaking in 25 mL of MOPS minimal media (Wilmes-Riesenberg & Wanner, 1992) with 2% glucose and chloramphenicol (35 μg/mL, if needed) in 250 mL baffled flasks. The overnight cultures were diluted to 0.05 optical density (OD) at 550 nm (OD_{550}) for the octanoic acid (C8) tolerance test or diluted to OD_{550} = 0.1 for testing membrane leakage, membrane fluidity, cell hydrophobicity, and cell membrane composition. Transformants were grown in LB media at 37ºC or 30ºC, with chloramphenicol (35 mg/L), ampicillin (100 mg/L), kanamycin (50 mg/L), or spectinomycin (50 mg/L) as needed.

Table 4.1 *Strains and plasmids used in this study.*

Plasmids or strains	Genotype or description	Reference
Plasmids		
pJMY-EEI82564	pTrc-EEI82564 thioesterase (TE10) from *Anaerococcus tetradius*, Amp^R	(Royce et al., 2015)
pJMY-Empty	pTrcHis B without the thioesterase (TE10), Amp^R	This study
Strains		
ML115	MG1655 (*ΔfabD, ΔpoxB, ΔackA-pta: cm*^{R})	(Li, Zhang, Agrawal, & San, 2012)
LAR1	ML115 evolved for C8 tolerance, Cm^R	(Royce et al., 2015)
LAR2	ML115 evolved for C8 tolerance, Cm^R	(Royce et al., 2015)
ML115+pJMY-Empty	ML115 with "empty" plasmid	This study
ML115+pJMY-EEI82564	ML115 with thioesterase for SCFA production	This study
LAR1+pJMY-Empty	LAR1 with "empty" plasmid	This study
LAR1+pJMY-EEI82564	LAR1 with thioesterase for SCFA production	This study
ML115+waaGInD	ML115 with insertion sequence removed from *waaG*	This study
ML115+waaGInD+rpoC*	ML115 with repaired *waaG* and *rpoC* mutation found in LAR1 and LAR2	This study
ML115+waaGInD+ΔbssS	ML115 with *waaG* repair and Δ*bssS*	This study

Growth analysis

Octanoic acid tolerance was determined by measuring OD$_{550}$ every hour. Overnight seed cultures were inoculated into 250 mL baffled flasks, which contained 25 mL MOPS with 2.0% (w/v) dextrose and 10 mM octanoic acid (1.44 g/L) with an initial media pH of 7.0 and an initial OD$_{550}$ of 0.05. The control groups used the same culture without the addition of octanoic acid. The flasks were incubated in a rotary shaker at 200 rpm and 37ºC. Cultures were taken and measured at OD$_{550}$ every hour.

RNA Isolation

Total RNA was isolated from saved cell pellets sampled at 6, 12, and 24 hours using RNeasy mini kit (Qiagen, Valencia, CA). Genomic DNA contamination was removed by Turbo DNA-free kit (Life Technologies, Carlsbad, CA), followed by the verification of total RNA using Agilent 2100 Bioanalyzer and RNA 600 Nano total RNA kit (Agilent, Santa Clara, CA). Next, ribosomal RNA (rRNA) was removed by Ribo-Zero (Bacteria) Magnetic kit (Illumina, San Diego, CA). Messenger RNA (mRNA) and other small RNA were purified and concentrated by Rneasy MinElute Cleanup kit (Qiagen, Valencia, CA). Agilent 2100 Bioanalyzer and RNA 6000 Pico mRNA kit (Agilent, Santa Clara, CA) were used to verify that the sample contained mRNA and other small RNA, but no rRNA. All procedures followed the manufacturers' user guide.

Fermentation for fatty acid production

The fatty acid production strains harboring the pJMY-EEI82564 plasmid by electroporation were grown on LB plates with ampicillin (100 mg/L) and incubated at 30ºC overnight. Individual colonies were precultured in 10 mL LB media with ampicillin (100 mg/L) in 250 mL flasks at 30ºC, 250 rpm with orbital shaking overnight. Seed cultures were then inoculated into 250 mL baffled flasks containing 50 mL of LB media with 1.5% dextrose, ampicillin (100 mg/L), and isopropyl-β-D-thiogalactopyranoside (IPTG) (1.0 mM) at an initial OD_{550} of 0.1. The flasks were incubated in a rotary shaker at 200 rpm and 30ºC, the culture samples were saved for testing fatty acid titer at 6, 12, 24, and 48 hours.

RNA-seq and short read analysis

For the transcriptomic analysis, RNA samples from the 6, 12, and 24 hour time points were obtained. These three time points were chosen to represent the lag phase, mid-log phase, and the stationary phase. After 24 hours, the OD_{550} value, the fatty acid titer, and the glucose concentration of the four strains did not vary significantly. The transcriptomic experiment performed was designed as a multi-factor experiment with three factor levels: strain, plasmid, and time. The two strains used in the experiment were the parent strain, ML115, and the evolved strain, LAR1. Plasmids were introduced into the strains to create a short chain fatty acid production condition (pJMY-EEI82564) and a control condition (pJMY-Empty). The plasmid pJMY-EEI82564 contains an acyl-acyl carrier protein thioesterase from *Anaerococcus tetradius* and an pJMY-Empty contains the same genetic material, but without the thioesterase gene. The multi-factor experiment had twelve combined factor levels with four biological replicates each for a total of forty-eight samples.

Single-end, directional RNA-Seq was performed by the Iowa State University DNA facility using the Illumina HiSeq 3000 platform with reads 100 base pairs (bp) in length. Reference-based assembly was performed using Rockhopper2 (version 2.0.3), which is designed specifically for bacterial systems (McClure et al., 2013). We created a reference transcriptome was corresponding to pJMY-EEI82564 that contained sequences of genes on the plasmid. Both *de novo* assembly and alignment to *E. coli* K-12 MG1655 (version U00096.3) (Blattner et al., 1997; Hayashi et al., 2006) reference transcriptome and pJMY-EEI82564 transcriptome were performed, which represented a combined 4,321 transcripts. Assembled RNA transcripts were used to further validate previously predicted genomic mutations.

Raw read counts from the Rockhopper2 assembly were analyzed with DESeq2 version 1.20.0 (Love, Huber, & Anders, 2014), a statistical package for differential expression analysis in R version 3.5.0.1 (R Core Team, 2018). DESeq2 was used to normalize raw count data and to test for differential expression between conditions and calculate pairwise $\log_2$ fold change (LFC). Differential expression analysis was performed in a pair-wise manner for strain contrasts at each time point and with each plasmid treatment. Differentially expressed genes were those with False Discovery Rate (FDR) adjusted p-value < 0.05.

Annotations from EcoCyc (Keseler et al., 2017) and RegulonDB (Gama-Castro et al., 2016) were used to identify genes with promoters associated with sigma factor 70 or belonging to the BasS-BasR regulon. Gene Ontology (GO) enrichment was performed to identify trends in gene function and localization annotations (Gene Ontology Consortium validation date 12/21/2015) (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). Gene lists were analyzed for overrepresented GO terms in the biological process, cellular component, and molecular function ontologies using BiNGO and a significance level of FDR corrected p-value < 0.05 (Maere, Heymans, & Kuiper, 2005).

E. coli Variant Analysis (EVA) software (Boggess et al., 2018) was used to generate a gene regulatory and metabolic network that reflected mutated features and downstream biological elements. Onto the network, we applied fold change data for strain contrasts to visualize altered transcript abundance for genes in the BasR regulon. We additionally modified EVA for use with candidate gene lists. Given a list of genes as input, we generated a gene regulatory and metabolic network to find potential interactions among perturbed genes.

Results and discussion

Fatty acid fermentation for ML115 and LAR1 with pJMY-EEI82564 or pJMY-Empty

We first compared the cell growth ability during short-chain fatty acid fermentation of *E. coli* ML115 and LAR1 with pJMY-EEI82564 and pJMY-Empty (Figure 4.1A). LAR1 with pJMY-EEI82564 reached the highest OD_{550} = 3.195 at 24 hours, then had a slightly decrease to OD_{550} = 2.828 at 48 h, finally increased to OD_{550} = 2.94 at 72 hours. ML115 with pJMY-EEI82564 reached the highest OD_{550} = 1.405 at 12 hours, then decreased to OD_{550} = 1.035 at 36 hours, finally increased to OD_{550} = 1.325 at 72 hours. LAR1 and ML115 with pJMY-Empty reached stationary phase (OD_{550} = 1.808 and 2.09) at 12 hours, then slight increased to final OD_{550}=2.218 and 2.475. The LAR1 strain with the pJMY-EEI82546 plasmid had the highest cell growth ability, even higher than LAR1 with the pJMY-Empty plasmid, demonstrating LAR1's improved growth rate when under SCFA production conditions.

From Figure 4.1**Error! Reference source not found.**B, we observed that the highest fatty acid titer of LAR1 with pJMY-EEI82564 could achieve 420.75 mg/L, which was 3-fold higher than that of ML115 with pJMY+EEI82564 (133.3 mg/L). The fatty acid titer of LAR1 and ML115 with pJMY-Empty was about 35 mg/L. The strains with pJMY-EEI82564 majorly produced free fatty acids (C4, C6, C8:0, C8:1, C10:0, C10:1, C12:1, C12:0, C14:1, C14:0, C16:1, C16:0, C18:1, and C18:0) during exponential phase, the C8 and C16 were the primary components (Figure 4.1E). The strains with pJMY-Empty majorly produced free fatty acids (C14:0, C16:1, C16:0, C18:1, and C18:0) during log phase (6 to 24 hours), the C16:0 and C18:0 were the primary components (Figure 4.1E).

LAR1 and ML115 strains with pJMY-EEI82564 consumed glucose primarily during the log phase (6 to 24 hours), and the glucose consumption rates were 0.35 g/L/h and 0.14 g/L/h, respectively (Figure 4.1C). In contrast, LAR1 and ML115 strains with pJMY-Empty consumed the most glucose during the lag and log phases (0 to 24 hours), the glucose consumption rate is 0.107 g/L/h and 0.165 g/L/h, respectively. After 24 hours, the glucose consumption ended for all four strains.

We also tested the pH of the fermentation media at different time points for all the strains (Figure 4.1D). Surprisingly, LAR1 with pJMY-EEI82564 was able to maintain a pH above 5.385, which had the highest short-chain fatty acid titer. The pH of the fermentation media of LAR1 with pJMY-Empty, LAR1 and ML115 with pJMY-EEI82564 was below 4.85 at 72 hours. When the LAR1 produced a large number of short-chain fatty acids, the LAR1 strain demonstrated a strategy to maintain the pH of media.

Identifying the *rpoC* mutation effect among differentially expressed genes

Differentially expressed genes were investigated for relation to the *rpoC* mutation by examining if they exhibited a consistent fold. We hypothesized that the transcriptomic signal caused by the *rpoC* mutation may be present under all time and plasmid conditions in strain contrasts. To search for genes affected by the *rpoC* mutation, we identified genes that exhibit a statistically significant and either consistently positive or negative fold change across all strain contrasts. Fifty-two genes had lower transcript abundance and sixty-five genes had higher transcript abundance in LAR1 versus ML115 for all time points and plasmid conditions. These sets of genes were filtered based on if they had a promoter that is associated with sigma factor 70, the primary sigma factor in *E. coli* K-12 MG1655 during exponential growth conditions (Jishage, Iwata, Ueda, & Ishihama, 1996).

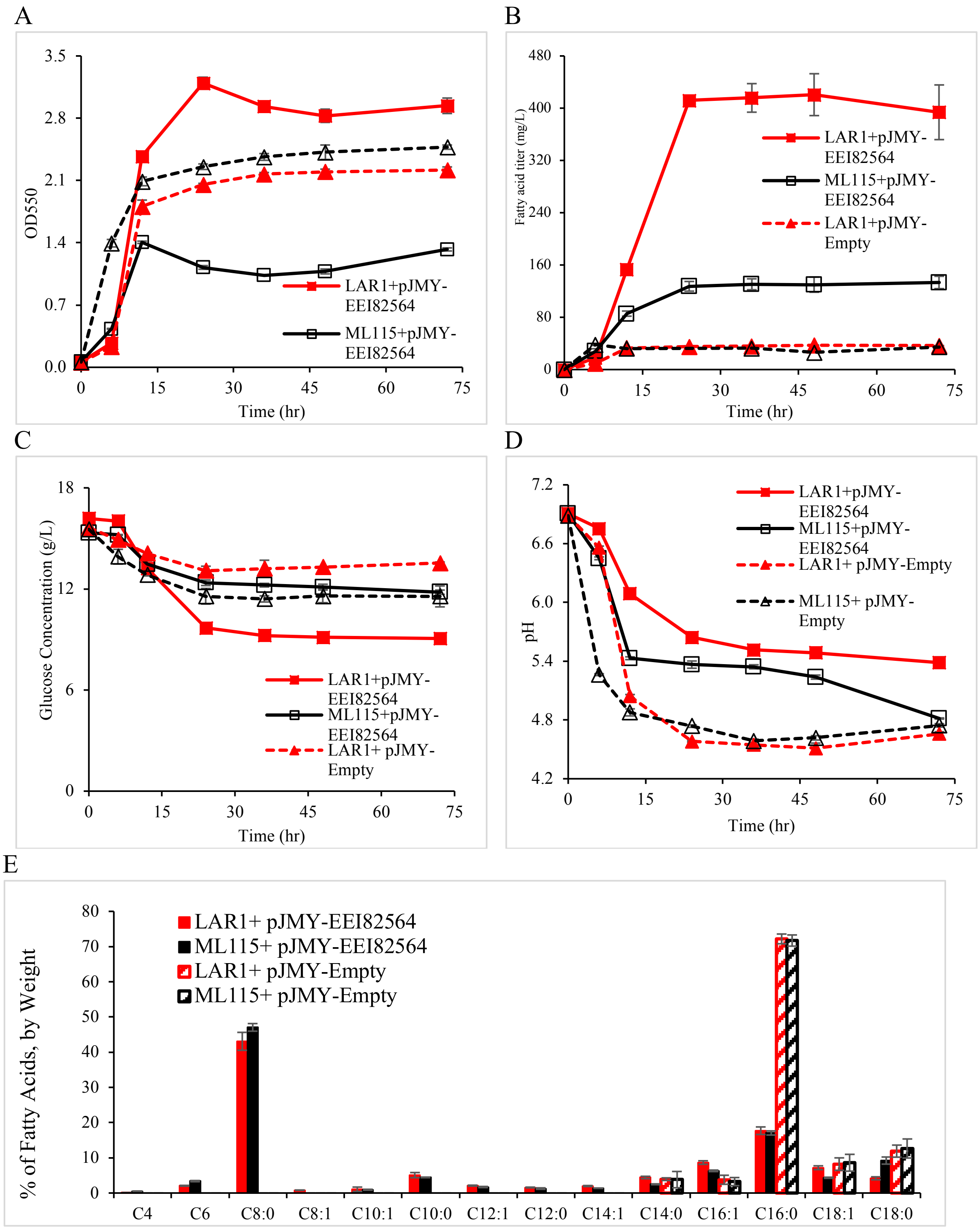


Figure 4.1 *Strain characteristics during fatty acid fermentation.*
A) The growth ability; B) Total fatty acid titer; C) Glucose concentration; D) pH of the media; E) Relative fatty acid distribution, by weight. Values are the average of four biological replicates, and error bars indicate standard deviation. Fermentation was performed with LB + 1.5% glucose, ampicillin (100 mg/L), and 1 mM IPTG at 30°C.

Eighteen downregulated and twenty upregulated genes were selected (Table 3.2). Similar filters were applied for other sigma factors and the alarmone, ppGpp (guanosine tetraphosphate and guanosine pentaphosphate), which binds to RNAP and regulates promoter selection.

A criteria of |log$_2$ fold change| > 0.5 for all strain contrasts was used to remove genes that exhibited small variations in transcript abundance. Thirteen genes had a log$_2$ fold change < -0.5 for all strain contrasts: *bssS*, *fliR*, *rcsA*, *dsrA*, *mntH*, *ssrA*, *speB*, *rplB*, *ugpQ*, *waaG*, *yjbE*, *phnP*, and *osmY*. Eleven genes exhibited a log$_2$ fold change > 0.5 in all strain contrasts: *hofC*, *hofB*, *btuF*, *mtn*, *fdnH*, *fdnI*, *pykA*, *nrdB*, *pka*, *yfiR*, *yfiN*, *yfiB*, *glyS*, *mtlD*, *fadA*, and *fadB*. The gene *waaG* was previously studied and contains an insertion sequence in the parent strain and is excluded from additional investigation into the effect of the *rpoC* mutation. Interestingly, the transcript abundance for *waaG* is greater in the parent strain with the non-functional copy of the gene. We attribute this to poor alignment in this region due to the insertion sequence.

We also considered that genes with sigma 70 promoters may be influenced by other mutations present in LAR1. We examined the gene regulatory network generated by EVA for the *basR* mutations and included the next three levels of gene regulation, which included 130 genes. When cross-referencing the list of genes downstream of these mutations with the differentially expressed genes mentioned above, we identify *osmY* and *fliR* as being indirectly regulated by the BasR transcription factor. In the case of *osmY*, BasR is a transcriptional activator for *csgD* (Ogasawara, Shinohara, Yamamoto, & Ishihama, 2012). CsgD represses transcription of *fliZ* (Dudin, Geiselmann, Ogasawara, Ishihama, & Lacour, 2014), and FliZ represses transcription of *csgD* and *osmY*. For *fliR (Pesavento et al., 2008;*

Pesavento & Hengge, 2012), BasR activates transcription of *qseB* (Guckes et al., 2013), QseB is believed to activate transcription of *flhC* and *flhD* (Sperandio, Torres, & Kaper, 2002), and FlhDC is a transcriptional activator for *fliR* (Brandi, Giangrossi, Giuliodori, & Falconi, 2016). In both cases, it is possible that the differences in transcript abundance between strains are indirectly affected by the *basR* mutation.

GO enrichment highlighted the relationship between *fdnH* and *fdnI* as components of the formate dehydrogenase complex but did not give insight into larger trends among the genes. Because these genes are in the same operon and co-transcribed, it is not surprising that they appeared in the same list. Other operons that appeared in our analysis include *yfiRNB*, *fadAB*, and *hofBC*. The *hofBC* genes have sequence similarity to protein secretion and fimbrial assembly genes (Whitchurch & Mattick, 1994) and *yfiRNB* genes is involved in exopolysaccharide biosynthesis in *Pseudomonas aeruginosa*. Because of our previous work characterizing the altered membrane and extracellular polymeric substances in LAR1, we believe these operons to be worthy of investigation. For these cases, we shall perform knockout experiments on the entire operon as a first test for relevance to evolved strain phenotype.

The *fadAB* operon is also of interest because of the genetic interventions introduced in the parent strain to modify the metabolism for improved fatty acid production. In ML115, a *fadD* knockout was added to deactivate the fatty acid beta-oxidation pathway (Figure 4.2). Two additional genetic interventions were made to inactivate acetate production: the deletion of *poxB* and *ack-pta* (Li et al., 2012).

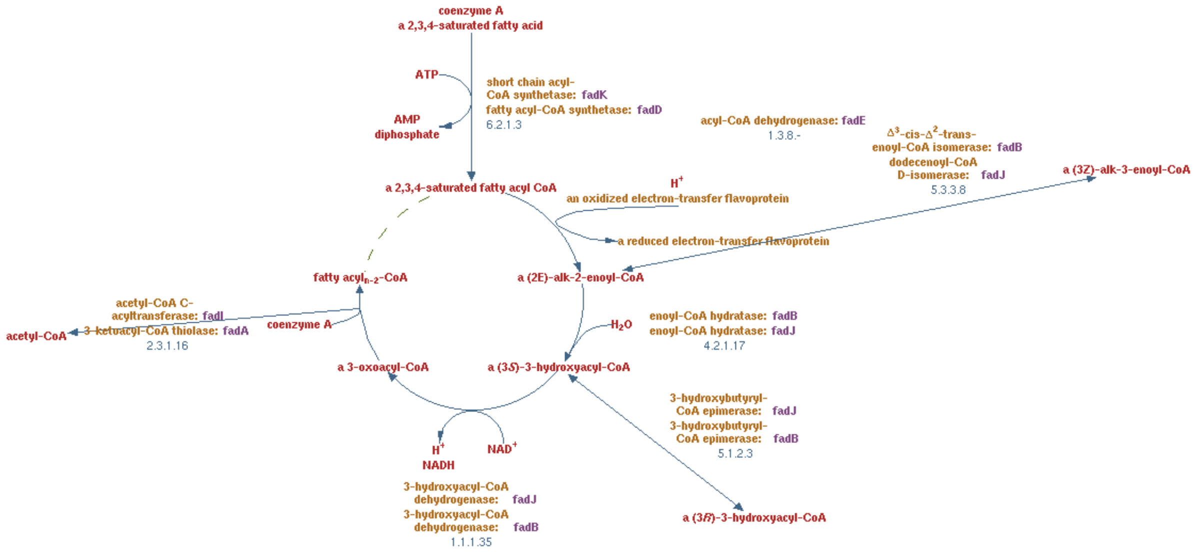

coenzyme A
a 2,3,4-saturated fatty acid
ATP
AMP
diphosphate
short chain acyl-CoA synthetase: fadK
fatty acyl-CoA synthetase: fadD
6.2.1.3
acyl-CoA dehydrogenase: fadE
1.3.8.-
Δ3-cis-Δ2-trans-enoyl-CoA isomerase: fadB
dodecenoyl-CoA D-isomerase: fadJ
5.3.3.8
a (3Z)-alk-3-enoyl-CoA
a 2,3,4-saturated fatty acyl CoA
H+
an oxidized electron-transfer flavoprotein
a reduced electron-transfer flavoprotein
fatty acyln-2-CoA
a (2E)-alk-2-enoyl-CoA
acetyl-CoA C-acyltransferase: fadI
3-ketoacyl-CoA thiolase: fadA
2.3.1.16
acetyl-CoA
coenzyme A
H2O
enoyl-CoA hydratase: fadB
enoyl-CoA hydratase: fadJ
4.2.1.17
a 3-oxoacyl-CoA
a (3S)-3-hydroxyacyl-CoA
3-hydroxybutyryl-CoA epimerase: fadJ
3-hydroxybutyryl-CoA epimerase: fadB
5.1.2.3
H+
NADH
NAD+
3-hydroxyacyl-CoA dehydrogenase: fadJ
3-hydroxyacyl-CoA dehydrogenase: fadB
1.1.1.35
a (3R)-3-hydroxyacyl-CoA

Figure 4.2 *Fatty acid beta oxidation pathway.*

The fatty acid beta oxidation pathway in E. coli (image retrieved from EcoCyc). Under aerobic conditions, FadD, FadB, and FadA break down fatty acids.

With fatty acyl-CoA synthetase (*fadD*) absent in ML115 and LAR1, it is surprising to see an increase in expression of *fadA* and *fadB* and even more interesting that there is greater transcript abundance for both genes in LAR1 when compared to ML115.

Other interesting genes from our list of candidates include *yjbE*, the most strongly downregulated gene for all LAR1-ML115 strain contrasts, for its role in biofilm formation and involvement in the production of extracellular polysaccharides (Ferrières, Aslam, Cooper, & Clarke, 2007). The gene *bssS* is also involved in biofilm formation and is the second most strongly downregulated gene. A *bssS* deletion has previously been reported as increasing biofilm formation and motility (Domka, Lee, & Wood, 2006).

Upon examining our candidate genes in the context of the gene regulatory and metabolic network generated by EVA, we find expected interactions that connect *fadA* and *fadB* as well as *fdnH* and *fdnI* (Figure 4.4A). Previously unexamined interactions are found between *pykA* and *mtlD* (Figure 4.4B). These genes are involved in the super pathway of glycolysis and Entner-Dourdoff. Both genes have a higher transcript abundance in LAR1 compared to ML115. A cluster containing *rcsA* and *yjbE* is also found in the network (Figure 4.4C). The gene *yjbE* was previously named as a gene of interest due to its role in extracellular polysaccharide production and because it exhibits the largest strain contrast. RcsAB is believed to activate transcription of *yjbE* (Ferrières et al., 2007) which would be consistent with a downregulation of *rcsA* and subsequent downregulation of *yjbE* as we see in LAR1 compared to ML115.

Expression of genes in the BasR regulon

Differentially expressed genes were compared with genes in the BasS-R regulon and up to three downstream levels of transcriptional regulation. We hypothesized that the production condition with the pJMY-EEI82564 plasmid would cause the most significant change in activity for the BasS-R two component signal transduction system, resulting in variation in expression among genes in the its regulon. To investigate genes in the BasR regulon, we used EVA to construct a gene regulatory and metabolic network that represented genes downstream of the transcription factor and included additional transcriptional regulation activities as previously described (Chen et al., 2018). Onto this network, we applied fold-change data for the strain contrast of interest (Figure 4.4) and BasR gene regulation activities. The variation between expected BasR transcription factor activity and observed direction in fold change for regulated genes speaks to the complexity of gene regulation and the fact that multiple regulators may affect gene expression. From our network analysis, we identified two operons of interest based on differential expression and gene function: *arnBCADTEF* and *csgDEFG*. Both operons will be knocked out and tested for effect on specific growth rate in 10 mM C8.

bssS **knockout improves specific growth rate at 10 mM C8**

The gene *bssS*, a regulator of biofilm, exhibited a consistent direction in fold change for all strain contrasts. In all cases, the transcript abundance of *bssS* was lower in LAR1 than ML115. Because of its reported role in biofilm formation and motility, it was selected for additional study. *bssS* has promoters associated with sigma 70 and sigma 32, the primary sigma factor for heat shock response. A knockout of *bssS* was introduced into the parent strain, ML115, with the *waaG* repair. We have already shown the profound effect damaging *waaG* has on phenotype but wish to examine if knocking out *bssS* contributes to octanoic acid tolerance. To test this, we measured growth rate of our reconstructed strains in 0 and 10 mM C8 (Figure 4.5A, B, and C). We find that specific growth rate is improved when *bssS* is knocked out (Figure 4.5D and E) compared with the parent strain with functional *waaG*, however this single intervention is not sufficient to recreate the phenotype observed with the variant *rpoC* (Figure 4.5B).

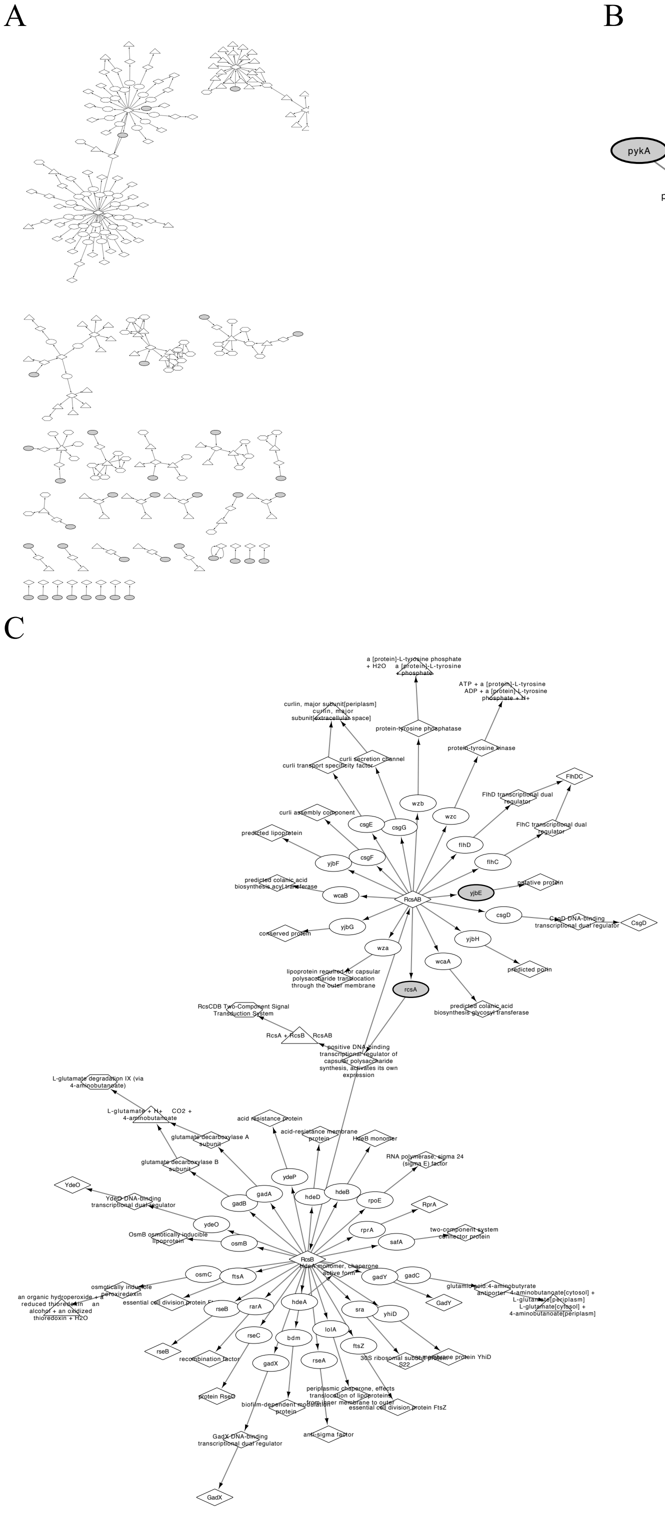

A
B
superpathway of glycolysis and Entner-Doudoroff
glycolysis I (from glucose 6-phosphate)
pykA
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
pyruvate kinase II
ADP + phosphoenolpyruvate + H+   pyruvate + ATP
glycolysis II (from fructose 6-phosphate)
mixed acid fermentation
superpathway of hexitol degradation (bacteria)
mannitol degradation I
D-mannitol 1-phosphate + NAD+   -D-fructofuranose 6-phosphate + NADH + H+
mannitol-1-phosphate 5-dehydrogenase
mtlD
C
a [protein]-L-tyrosine phosphate + H2O   a [protein]-L-tyrosine + phosphate
ATP + a [protein]-L-tyrosine   ADP + a [protein]-L-tyrosine phosphate + H+
curlin, major subunit[periplasm]   curlin, major subunit[extracellular space]
protein-tyrosine phosphatase
protein-tyrosine kinase
curli secretion channel
curli transport specificity factor
FlhDC
FlhD transcriptional dual regulator
wzb
curli assembly component
csgE
csgG
wzc
FlhC transcriptional dual regulator
predicted lipoprotein
flhD
csgF
yjbF
flhC
predicted colanic acid biosynthesis acyl transferase
putative protein
wcaB
RcsAB
yjbE
csgD
CsgD DNA-binding transcriptional dual regulator
CsgD
yjbG
conserved protein
yjbH
wza
wcaA
predicted porin
lipoprotein required for capsular polysaccharide translocation through the outer membrane
rcsA
predicted colanic acid biosynthesis glycosyl transferase
RcsCDB Two-Component Signal Transduction System
RcsA + RcsB   RcsAB
positive DNA-binding transcriptional regulator of capsular polysaccharide synthesis, activates its own expression
L-glutamate degradation IX (via 4-aminobutanoate)
L-glutamate + H+   CO2 + 4-aminobutanoate
acid resistance protein
acid-resistance membrane protein
HdeB monomer
glutamate decarboxylase A subunit
RNA polymerase, sigma 24 (sigma E) factor
glutamate decarboxylase B subunit
ydeP
YdeO
gadA
hdeD
hdeB
rpoE
YdeO DNA-binding transcriptional dual regulator
gadB
RprA
ydeO
rprA
two-component system connector protein
OsmB osmotically inducible lipoprotein
osmB
safA
RcsB
HdeA monomer, chaperone active form
osmC
ftsA
gadY
gadC
osmotically inducible peroxiredoxin
glutamic acid:4-aminobutyrate antiporter   4-aminobutanoate[cytosol] + L-glutamate[periplasm]   L-glutamate[cytosol] + 4-aminobutanoate[periplasm]
an organic hydroperoxide + a reduced thioredoxin   an alcohol + an oxidized thioredoxin + H2O
essential cell division protein FtsA
rseB
rarA
hdeA
sra
yhiD
GadY
rseC
bdm
lolA
ftsZ
rseB
recombination factor
gadX
rseA
30S ribosomal subunit protein S22
inner membrane protein YhiD
protein RseC
periplasmic chaperone, effects translocation of lipoproteins from inner membrane to outer
biofilm-dependent modulation protein
essential cell division protein FtsZ
anti-sigma factor
GadX DNA-binding transcriptional dual regulator
GadX

Figure 4.3 *EVA-generated gene regulatory and metabolic networks for genes in Table 3.2.*

A. The full network for all genes of interest: 283 nodes and 283 edges; B. Selected cluster from (A) with pykA and mtlD genes; C. The largest cluster in the network contains regulatory interactions that connect rcsA and yjbE.
]

Table 4.2 *Genes with consistent and statistically significant fold changes for all strain contrasts.*

Genes are ordered by the smallest log$_2$ *fold change (LFC) among all strain contrasts, ascending, with a divider separating genes that had a lower transcript abundance in LAR1 with genes that had a higher transcript abundance in LAR1.*

b-num	Name	Product	LAR1-ML115 (min LFC)
b4026	*yjbE*	extracellular polysaccharide production threonine-rich protein	-3.02
b1060	*bssS*	biofilm regulator	-1.78
b3449	*ugpQ*	glycerophosphodiester phosphodiesterase, cytosolic	-1.38
b1954	*dsrA*	small regulatory RNA	-1.32
b4376	*osmY*	periplasmic protein	-1.23
b2621	*ssrA*	tmRNA	-1.21
b3631	*waaG*	glucosyltransferase I	-1.20
b2392	*mntH*	manganese/divalent cation transporter	-0.99
b1951	*rcsA*	transcriptional regulator of colanic acid capsular biosynthesis	-0.89
b1950	*fliR*	flagellar export pore protein	-0.89
b4092	*phnP*	5-phospho-alpha-D-ribosyl 1,2-cyclic phosphate phosphodiesterase	-0.67
b3317	*rplB*	50S ribosomal subunit protein L2	-0.58
b2937	*speB*	agmatinase	-0.52
b0244	*thrW*	Thr tRNA	-0.47
b3924	*fpr*	ferredoxin-NADP reductase	-0.44
b0880	*cspD*	inhibitor of DNA replication, cold shock protein homolog	-0.43
b0143	*pcnB*	poly(A) polymerase	-0.41
b0345	*lacI*	lactose-inducible lac operon transcriptional repressor	-0.33
b2558	*mltF*	membrane-bound lytic transglycosylase F, murein hydrolase	0.29
b2509	*xseA*	exonuclease VII, large subunit	0.41
b2686	*emrB*	multidrug efflux system protein	0.42
b2913	*serA*	D-3-phosphoglycerate dehydrogenase	0.43
b1854	*pykA*	pyruvate kinase II	0.51
b2603	*yfiR*	putative periplasmic inhibitor of YfiN activity	0.69
b3600	*mtlD*	mannitol-1-phosphate dehydrogenase, NAD-dependent	0.75
b2235	*nrdB*	ribonucleoside-diphosphate reductase 1, beta subunit, ferritin-like protein	0.76
b3845	*fadA*	3-ketoacyl-CoA thiolase (thiolase I)	0.76

Table 4.2 *(continued)*

b-num	Name	Product	LAR1-ML115 (min LFC)
b3559	*glyS*	glycine tRNA synthetase, beta subunit	0.90
b0106	*hofC*	assembly protein in type IV pilin biogenesis, transmembrane protein	0.91
b0159	*mtn*	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	0.93
b0107	*hofB*	T2SE secretion family protein; P-loop ATPase superfamily protein	0.94
b1476	*fdnI*	formate dehydrogenase-N, cytochrome B556 (gamma) subunit, nitrate-inducible	0.97
b2584	*pka*	protein lysine acetyltransferase	1.12
b3846	*fadB*	fused 3-hydroxybutyryl-CoA epimerase/delta(3)-cis-delta(2)-trans-enoyl-CoA isomerase/enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase	1.16
b2605	*yfiB*	OM lipoprotein putative positive effector of YfiN activity	1.24
b0158	*btuF*	vitamin B12 transporter subunit: periplasmic-binding component of ABC superfamily	1.35
b1475	*fdnH*	formate dehydrogenase-N, Fe-S (beta) subunit, nitrate-inducible	1.40
b2604	*yfiN*	putative membrane-anchored diguanylate cyclase with an N-terminal periplasmic domain	1.43

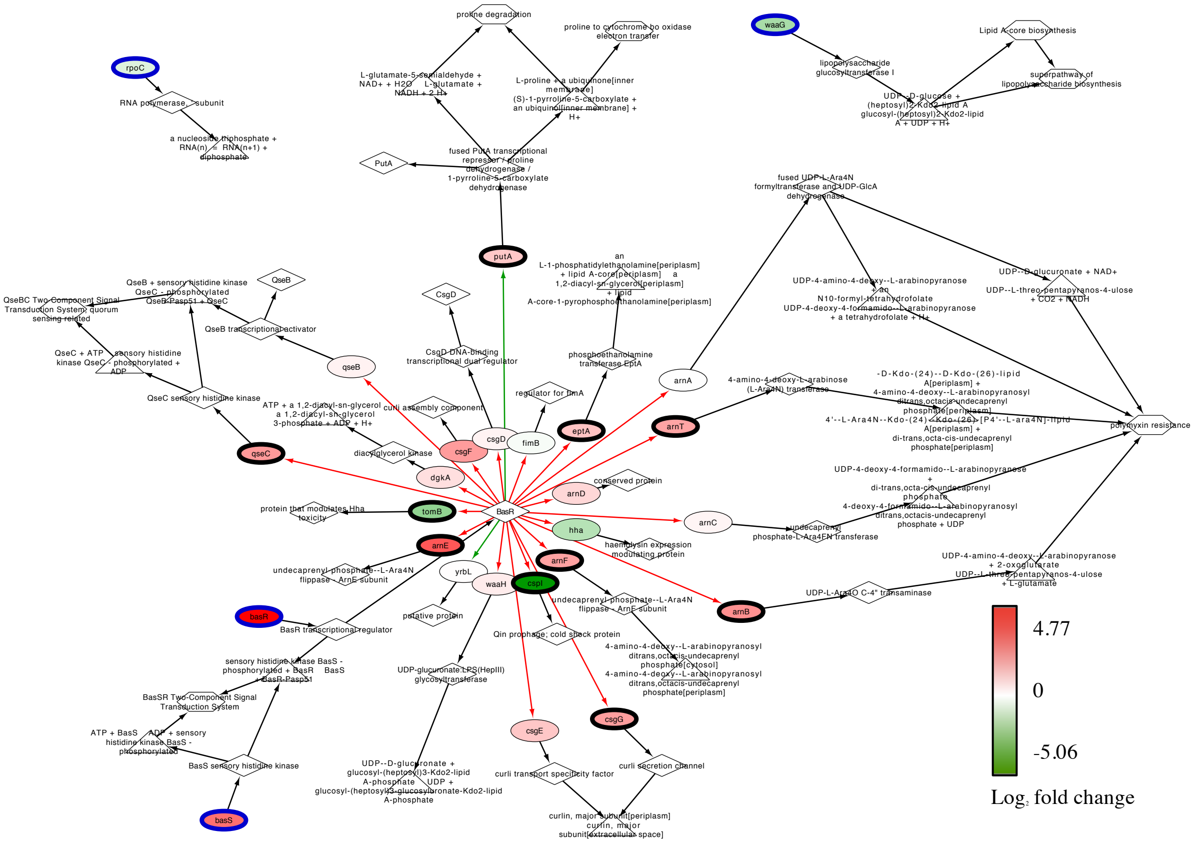

rpoC
RNA polymerase, ' subunit
a nucleoside triphosphate + RNA(n) = RNA(n+1) + diphosphate
proline degradation
proline to cytochrome bo oxidase electron transfer
L-glutamate-5-semialdehyde + NAD+ + H2O L-glutamate + NADH + 2 H+
L-proline + a ubiquinone[inner membrane] (S)-1-pyrroline-5-carboxylate + an ubiquinol[inner membrane] + H+
PutA
fused PutA transcriptional repressor / proline dehydrogenase / 1-pyrroline-5-carboxylate dehydrogenase
putA
waaG
lipopolysaccharide glucosyltransferase I
UDP--D-glucose + (heptosyl)2-Kdo2-lipid A glucosyl-(heptosyl)2-Kdo2-lipid A + UDP + H+
Lipid A-core biosynthesis
superpathway of lipopolysaccharide biosynthesis
fused UDP-L-Ara4N formyltransferase and UDP-GlcA dehydrogenase
UDP-4-amino-4-deoxy--L-arabinopyranose + an N10-formyl-tetrahydrofolate UDP-4-deoxy-4-formamido--L-arabinopyranose + a tetrahydrofolate + H+
UDP--D-glucuronate + NAD+ UDP--L-threo-pentapyranos-4-ulose + CO2 + NADH
an L-1-phosphatidylethanolamine[periplasm] + lipid A-core[periplasm] a 1,2-diacyl-sn-glycerol[periplasm] + lipid A-core-1-pyrophosphoethanolamine[periplasm]
QseB + sensory histidine kinase QseC - phosphorylated QseB-Pasp51 + QseC
QseB
QseBC Two-Component Signal Transduction System; quorum sensing related
QseB transcriptional activator
CsgD
CsgD DNA-binding transcriptional dual regulator
phosphoethanolamine transferase EptA
QseC + ATP sensory histidine kinase QseC - phosphorylated + ADP
qseB
arnA
4-amino-4-deoxy-L-arabinose (L-Ara4N) transferase
-D-Kdo-(24)--D-Kdo-(26)-lipid A[periplasm] + 4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[periplasm] 4'--L-Ara4N--Kdo-(24)--Kdo-(26)-[P4'--L-ara4N]-lipid A[periplasm] + di-trans,octa-cis-undecaprenyl phosphate[periplasm]
QseC sensory histidine kinase
ATP + a 1,2-diacyl-sn-glycerol a 1,2-diacyl-sn-glycerol 3-phosphate + ADP + H+
curli assembly component
regulator for fimA
csgD
fimB
eptA
arnT
polymyxin resistance
csgF
qseC
diacylglycerol kinase
dgkA
conserved protein
arnD
UDP-4-deoxy-4-formamido--L-arabinopyranose + di-trans,octa-cis-undecaprenyl phosphate 4-deoxy-4-formamido--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate + UDP
protein that modulates Hha toxicity
tomB
BasR
arnC
undecaprenyl phosphate-L-Ara4FN transferase
hha
haemolysin expression modulating protein
arnE
UDP-4-amino-4-deoxy--L-arabinopyranose + 2-oxoglutarate UDP--L-threo-pentapyranos-4-ulose + L-glutamate
undecaprenyl-phosphate--L-Ara4N flippase - ArnE subunit
yrbL
arnF
waaH
cspI
undecaprenyl-phosphate--L-Ara4N flippase - ArnF subunit
UDP-L-Ara4O C-4" transaminase
arnB
basR
putative protein
BasR transcriptional regulator
Qin prophage; cold shock protein
4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[cytosol] 4-amino-4-deoxy--L-arabinopyranosyl ditrans,octacis-undecaprenyl phosphate[periplasm]
sensory histidine kinase BasS - phosphorylated + BasR BasS + BasR-Pasp51
UDP-glucuronate:LPS(HepIII) glycosyltransferase
BasSR Two-Component Signal Transduction System
ATP + BasS ADP + sensory histidine kinase BasS - phosphorylated
csgG
csgE
curli secretion channel
BasS sensory histidine kinase
UDP--D-glucuronate + glucosyl-(heptosyl)3-Kdo2-lipid A-phosphate UDP + glucosyl-(heptosyl)3-glucosyluronate-Kdo2-lipid A-phosphate
curli transport specificity factor
curlin, major subunit[periplasm] curlin, major subunit[extracellular space]
basS
4.77
0
-5.06
Log₂ fold change

Figure 4.4 *EVA-generated network with RNA-seq data.*

One level of gene regulation represented using mutated features as seed nodes in the network. Blue node borders correspond to mutated features. Colored interaction arrows show gene regulation; red = activation, green = repression. Node coloring depicts the strain contrast for the production condition during the exponential growth phase: log_2 fold change for LAR1+pJMY-EEI82564 – ML115+pJMY-EEI82564 at 12 hours. Emphasized node borders indicate p-value < 0.05.

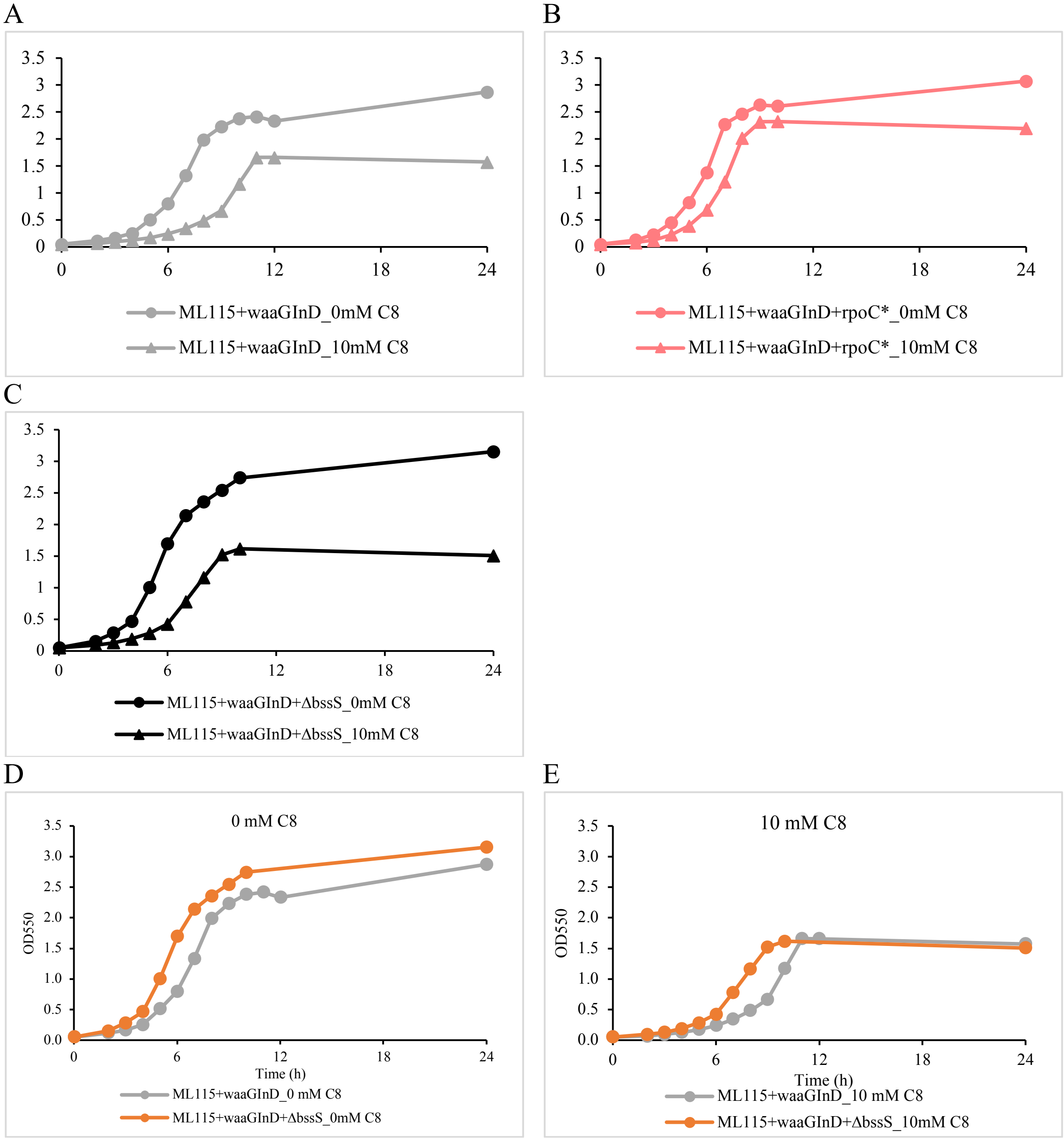


Figure 4.5 *The effect on specific growth rate of ΔbssS in 0 and 10 mM C8.*

A. growth rate for the parent strain with repaired waaG; B. growth rate for the parent strain with repaired waaG and rpoC mutation; C. growth rate for the parent strain with repaired waaG and bssS knockout; D. the effect of the bssS knockout on growth in 0 mM C8; E. the effect of the bssS knockout on growth in 10 mM C8.

Conclusions

In this work we have performed RNA-seq experiments for the parent and evolved strains and identified genes that are consistently differentially expressed in different growth phases and in control and fatty acid production conditions. We additionally examined genes with significant strain contrasts and selected candidates for further study. We modified EVA to construct a gene regulatory and metabolic network using this set of genes as seeds. From this network, we were able to find additional connections through gene regulation among genes of interest. The network also highlighted metabolic pathways that could be affected by differentially expressed genes. We also visualized fold-change data on the EVA network for the BasR regulon.

We have begun constructing gene knockout strains based on our list of candidate genes to test their effect on growth with exogenous C8. Thus far, Δ*bssS* has been tested and found to improve specific growth rate but does not account for the improvement we see from the *rpoC* mutation. The effects of the variant *basR* and *rpoC* genes on transcription may be broad and could involve multiple genes and pathways. We have presented our bioinformatics-based approach to identifying gene candidates for further investigation based on relationship to mutated features, differential expression in transcriptomic analysis, and gene regulation and metabolic network relationships.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1), 25-29. doi:10.1038/75556

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. Science, 277(5331), 1453-1462. doi:papers3://publication/uuid/5530EBDB-8A2B-4B0F-8B27-8116B7DB359F

Boggess, E., Jarboe, L., & Dickerson, J. (2018). Mutation analysis for metabolic evolution experiments in Escherichia coli. BMC Bioinformatics, Submitted.

Brandi, A., Giangrossi, M., Giuliodori, A. M., & Falconi, M. (2016). An Interplay among FIS, H-NS, and Guanosine Tetraphosphate Modulates Transcription of the Escherichia coli cspA Gene under Physiological Growth Conditions. Front Mol Biosci, 3, 19. doi:10.3389/fmolb.2016.00019

Chen, Y., Boggess, E., Dickerson, J., & Jarboe, L. (2018). Genome-level reverse engineering of Escherichia coli evolved for increased short-chain fatty acid tolerance and production. Metabolic Engineering, To be submitted

Domka, J., Lee, J., & Wood, T. K. (2006). YliH (BssR) and YceP (BssS) regulate Escherichia coli K-12 biofilm formation by influencing cell signaling. Appl Environ Microbiol, 72(4), 2449-2459. doi:10.1128/AEM.72.4.2449-2459.2006

Dudin, O., Geiselmann, J., Ogasawara, H., Ishihama, A., & Lacour, S. (2014). Repression of flagellar genes in exponential phase by CsgD and CpxR, two crucial modulators of Escherichia coli biofilm formation. J Bacteriol, 196(3), 707-715. doi:10.1128/JB.00938-13

Ferrières, L., Aslam, S. N., Cooper, R. M., & Clarke, D. J. (2007). The yjbEFGH locus in Escherichia coli K-12 is an operon encoding proteins involved in exopolysaccharide production. Microbiology, 153(Pt 4), 1070-1080. doi:10.1099/mic.0.2006/002907-0

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., . . . Collado-Vides, J. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res, 44(D1), D133-143. doi:10.1093/nar/gkv1156

Guckes, K. R., Kostakioti, M., Breland, E. J., Gu, A. P., Shaffer, C. L., Martinez, C. R., . . . Hadjifrangiskou, M. (2013). Strong cross-system interactions drive the activation of the QseB response regulator in the absence of its cognate sensor. Proc Natl Acad Sci U S A, 110(41), 16592-16597. doi:10.1073/pnas.1315320110

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., . . . Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol Syst Biol, 2, 2006.0007. doi:10.1038/msb4100049

Jishage, M., Iwata, A., Ueda, S., & Ishihama, A. (1996). Regulation of RNA polymerase sigma subunit synthesis in Escherichia coli: intracellular levels of four species of sigma subunit under various growth conditions. J Bacteriol, 178(18), 5447-5451.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., . . . Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res, 45(D1), D543-D550. doi:10.1093/nar/gkw1003

Li, M., Zhang, X., Agrawal, A., & San, K. Y. (2012). Effect of acetate formation pathway and long chain fatty acid CoA-ligase on the free fatty acid production in E. coli expressing acy-ACP thioesterase from Ricinus communis. Metab Eng, 14(4), 380-387. doi:10.1016/j.ymben.2012.03.007

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15(12), 550. doi:10.1186/s13059-014-0550-8

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21(16), 3448-3449. doi:papers3://publication/doi/10.1093/bioinformatics/bti551

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., . . . Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. Nucleic acids research, 41(14), e140-e140. doi:papers3://publication/doi/10.1093/nar/gkt444

Ogasawara, H., Shinohara, S., Yamamoto, K., & Ishihama, A. (2012). Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. Microbiology, 158(Pt 6), 1482-1492. doi:10.1099/mic.0.057745-0

Pesavento, C., Becker, G., Sommerfeldt, N., Possling, A., Tschowri, N., Mehlis, A., & Hengge, R. (2008). Inverse regulatory coordination of motility and curli-mediated adhesion in Escherichia coli. Genes Dev, 22(17), 2434-2446. doi:10.1101/gad.475808

Pesavento, C., & Hengge, R. (2012). The global repressor FliZ antagonizes gene expression by σS-containing RNA polymerase due to overlapping DNA binding specificity. Nucleic Acids Res, 40(11), 4783-4793. doi:10.1093/nar/gks055

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.4.1). Vienna, Austria: R Foundation for Statistical Computing.

Royce, L. A., Yoon, J. M., Chen, Y., Rickenbach, E., Shanks, J. V., & Jarboe, L. R. (2015). Evolution for exogenous octanoic acid tolerance improves carboxylic acid production and membrane integrity. Metab Eng, 29, 180-188. doi:10.1016/j.ymben.2015.03.014

Sperandio, V., Torres, A. G., & Kaper, J. B. (2002). Quorum sensing Escherichia coli regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in E. coli. Mol Microbiol, 43(3), 809-821.

The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res, 45(D1), D331-D338. doi:10.1093/nar/gkw1108

Whitchurch, C. B., & Mattick, J. S. (1994). Escherichia coli contains a set of genes homologous to those involved in protein secretion, DNA uptake and the assembly of type-4 fimbriae in other bacteria. Gene, 150(1), 9-15.

Wilmes-Riesenberg, M. R., & Wanner, B. L. (1992). TnphoA and TnphoA' elements for making and switching fusions for study of transcription, translation, and cell surface localization. Journal of bacteriology, 174(14), 4558-4575. doi:papers3://publication/doi/10.1128/jb.174.14.4558-4575.1992

CHAPTER 5. SUMMARY, FUTURE WORK, AND CONCLUSIONS

Microbial evolution as a strategy for strain engineering has been shown to be an effective tool for enhancing tolerance of *E. coli* to carboxylic acids for greater yields, titers, and productivity. We also examined case studies that employed this strategy for studying antibiotic resistance in bacteria. In both cases, continuous fermentation under selective pressure is used to obtain strains that exhibit a desired phenotype. Such evolution studies may be performed iteratively and can also include targeted genetic interventions. In all cases, the metabolic evolution itself is a black box technique and the mutations acquired during the experiment are not known until genomes of evolved strains are sequenced. Reverse engineering of evolved strains necessary to identify genetic variations from the parent strain and discern which mutations are relevant the evolved phenotype and by what mechanisms.

The task of reverse engineering evolved strains remains a significant challenge. Inconsistencies among short read assembly software, sequencing errors, and complex mutations such as rearrangements and large indels can lead to inaccurate mutation predictions. Verification of predicted mutations by PCR and Sanger sequencing is necessary to have confidence in genome annotations.

This work has focused primarily on characterizing mutations identified in comparative genomic analysis with the goal of relating genotype to phenotype, a significant challenge in microbial engineering. Chapter 2 present the software developed to work toward this goal and support reverse engineering efforts. We present a strategy that is appropriate for many types of mutations, but could be extended as analysis methods advance. While we built our software for *E. coli*, the same concepts apply to other bacterial systems. To implement a version of EVA for a different organism, a complete genome, gene models, and sufficient gene regulation and metabolic pathway data are required. Organisms without well-characterized gene regulation and metabolism could still benefit from the annotation and sequence analysis components of EVA. Even for *E. coli*, we must rely on incomplete models of gene regulation and cellular metabolism, and so our results are also incomplete. The best publicly available databases for this model organism may not include some published regulatory links, such as was the case for the BasR regulon. Other regulation links may only be from computation predictions based on sequence similarity to known binding sites.

There is the additional challenge of presenting meaningful biological network information without including too much extraneous information. To provide options for biologists, EVA generates biological networks in three ways: a complete network with all available regulatory links, a mutation-interaction network with all nodes and edges reachable by two or more mutations, and a shortest-path network with minimal paths between mutated features. Additional network analysis and statistics could be derived from these networks, an analysis that could be increasingly valuable as the number of mutations in an experiment grows or in cases where additional regulatory steps are included in the network.

In Chapter 3, we used the EVA software to aid in mutation interpretation for *E. coli* evolved for improved octanoic acid tolerance. In Chapter 4, we continued our reverse engineering efforts with an associated RNA-seq study. The utilization of transcriptomic and other omics data in reverse engineering microbial strains provides valuable information about cellular activities which cannot be known from genomics studies alone. The gene regulatory and metabolic networks generated by EVA and using mutated features as seed nodes are readily integrated with transcriptomics data. In addition, we modified EVA to generate new networks from differentially expressed genes and were able to identify relationships among these genes that may point to the underlying mechanisms responsible for phenotype.

It would be interesting to incorporate other types of omics data (e.g., proteomics, flux analysis) into our analysis in the future. Additionally, using EVA networks to build simple models for gene activation and repression could aid in interpretation and potentially identify missing regulatory links. And finally, utilizing EVA for rational engineering could provide researchers with multiple targets for incorporating a desired genomic intervention, such as promoter or transcription factor binding site modification.

APPENDIX A. IDENTIFICATION OF MUTATIONS IN EVOLVED BACTERIAL GENOMES

Methods book chapter in: Hal S. Alper (ed.), Systems Metabolic Engineering: Methods and Protocols, Methods in Molecular Biology, vol. 985:249-267

Liam Royce, Erin Boggess, Tao Jin, Julie Dickerson, Laura Jarboe

Summary

Directed laboratory evolution is a common technique to obtain an evolved bacteria strain with a desired phenotype. This technique is especially useful as a supplement to rational engineering for complex phenotypes such as increased biocatalyst tolerance to toxic compounds. However, reverse engineering efforts are required in order to identify the mutations that occurred, including single polymorphisms (SNPs), insertions/deletions (indels), duplications, and rearrangements. In this protocol, we describe the steps to 1) obtain and sequence the genomic DNA 2) process and analyze the genomic DNA sequence data, and 3) verify the mutations by Sanger resequencing.

Introduction

Bacteria acting as biocatalysts for production of biorenewable fuels and chemicals are often faced with product-mediated inhibition. For example, ethanol was shown to negatively impact growth and structure of *E. coli* and yeast (Ingram & Buttke, 1984; Trinh, Huffer, Clark, Blanch, & Clark, 2010); the effects of succinate were revealed on the membrane and enzymes of yeast (Duro & Serrano, 1981; Smith, Janknecht, & Maher, 2007); butanol was shown to inhibit the growth and sugar uptake rate of Clostridium acetobutylicum (Schwarz, Kuit, Grimmler, Ehrenreich, & Kengen, 2012; Winkler & Kao, 2011).

Lignocellulosic biomass has been extensively utilized as a source of carbon and energy for the fermentative production of ethanol and other biorenewable fuels (Jarboe, Grabar, Yomano, Shanmugan, & Ingram, 2007; Jørgensen, Kristensen, & Felby, 2007; C. Li, Qian, & Zhao, 2008). However, the sugar streams released from this biomass frequently contain inhibitory contaminants that inhibit the growth and substrate utilization of microorganism (Miller, Jarboe, Turner, et al., 2009; Zaldivar & Ingram, 1999; Zaldivar, Martinez, & Ingram, 1999).

Thus, the fermentative production of biorenewable fuels and chemicals is associated with both inhibitory contaminants in the feedstock and inhibitory products; in these cases, it can be sometimes useful to increase the tolerance of the biocatalyst to these inhibitory compounds. Metabolic evolution is frequently used to increase the tolerance of bacteria to inhibitory compounds. Directed evolution is when researchers can enhance desired features, such as increased tolerance of inhibitory compounds, by selecting for random mutations under appropriate selective pressure. While metabolic evolution is sufficient to acquire a strain with the desired phenotype, it is often of interest to identify the mutations acquired during the evolutionary process.

Reverse engineering can yield a roadmap for reproducing the desired phenotype or behavior in other biocatalysts. This method begins with whole-genome sequencing using high-throughput sequencing technology, such as Illumina's sequencing by synthesis technique. Bioinformatics methods known as de novo assembly and mapping (or alignment) are used to analyze the short read data and reconstruct the genome (Langmead, Trapnell, Pop, & Salzberg, 2009; H. Li & Durbin, 2009; H. Li et al., 2009). By obtaining DNA sequences of the parent and evolved organism genomes, it is possible to perform a comparative analysis and identify variations in the evolved strain. Isolation of bacteria genomes is a standard procedure. Sequencing platforms are changing rapidly in their throughput and chemistry to increase availability and fidelity of sequence data (Metzker, 2010). As sequencing data becomes more readily available, there are many challenges to the processing and analysis of sequence data, which is costly and time consuming. Thus, automation by programs alleviates the burden of manual analysis. The finishing step and gap filling in DNA sequence analysis is the bottleneck in automation (D. Gordon, Abajian, & Green, 1998). In the recent decade, there has been a great amount of improvement in automating the process with computer programs; however, this step still requires human intervention.

As the genotype of the evolved strain is defined, hypotheses are formed regarding the roles of mutations in the context of the phenotype. As researchers elucidate which mutations improve fitness, the intent is to infer the mechanisms that lead to the increased tolerance to toxicity and then proceed with rational engineering techniques (Jarboe et al., 2007; Miller, Jarboe, Yomano, et al., 2009). This can also enable to identify the function of the undercharacterized enzymes and pathways (Jarboe, 2011). However, the focus of this chapter is to describe the use of genome sequence analysis to identify the mutations acquired in an evolved strain. Determination of which of these mutations impact the phenotype and understanding the mechanism of the mutation’s function is outside the scope of this chapter.

Materials

All materials used are standard kits and reagents. Software for high throughput

sequence analysis generally requires UNIX/Linux operating systems with a large amount of memory and storage. Both free and commercial software packages are available for analyzing high-throughput sequencing data. All software included in this protocol is open source unless otherwise noted.

Genomic DNA Purification and Sequencing

1. Lauria Broth (LB) for growing bacteria cells: dissolve at 25 g/L in nanopure water and filter-sterilize using a 0.22 CA bottle top filter.
2. 1.5mL microfuge tubes and 50mL centrifuge tubes for sample processing.
3. QIAGEN DNeasy® Blood & Tissue kit for genomic DNA isolation and purification. Buy RNase A and 100 % ethanol separately.
4. Accublock Digital Dry Bath for temperature-controlled incubation.
5. NanoDrop Spectrophotometer for genomic DNA quantification and quality control.
6. Illumina cBot System and Illumina TruSeq PE Cluster Kit -GA for cluster generation, Illumina GAII sequencing instrument for short-read whole genome sequencing (available at a university core facility, prices vary).

Bioinformatics Software for High-Throughput Sequence Data

1. Galaxy is a scientific workflow system for high-throughput sequence data preprocessing, integration, and analysis. A free public server is available, but most users will need to download and install the open source Galaxy software locally due to the upload limitations and to preserve data privacy. UNIX/Linux and Mac OS X are supported and a recent version of Python must be installed (Blankenberg et al., 2010; Giardine et al., 2005; Goecks, Nekrutenko, Taylor, & Team, 2010).

2. FastQC provides quality control checks for raw sequence data and generates summary graphs and basic statistics. FastQC is available through the Galaxy interface or for download and independent installation (Andrews, 2012).
3. FASTX-Toolkit is a collection of scripts for manipulating raw sequence data. It includes conversion, trimming, and filtering tools and will generate some quality statistics. The FASTX-Toolkit is distributed with Galaxy or can be downloaded and installed independently (A. Gordon).
4. Mapping software: Bowtie, Bowtie 2, and BWA are popular short read aligners that distributed under the GPLv3 license. Bowtie and BWA are distributed with Galaxy. Memory requirements vary by algorithm and input data, but at least 2GB memory required and at least 4GB is recommended. Multiple processors can also improve alignment speed. It is critical to read the manual for mapping software because different parameters will generate different alignments.
5. *de novo* assembly software: Velvet and ABySS (available for download and distributed under the GPLv3 license) are examples the many available de Buijn graph-based assemblers (Simpson et al., 2009; Zerbino & Birney, 2008). Other assemblers that use an overlap/layout/consensus approach are available, but take considerably longer to assemble short reads and are not considered for this protocol. While many assemblers support 32-bit platforms, a 64-bit machine is recommended and memory requirements vary by algorithm, short read data, and selected k-mer length. It is critical to read the assembler manuals because different parameters generate different contigs/scaffolds.

6. Basic Local Alignment Search Tool (BLAST) is the most widely used sequence similarity tool. A web interface is available through NCBI, but a local installation of the BLAST+ open source applications provide a command line usage (Camacho et al., 2009).
7. SAMtools is a collection of utilities for manipulating alignments. BCFtools, which is distributed with SAMtools, performs variant calling. SAMtools is distributed with Galaxy and can also be independently installed (H. Li & Durbin, 2009; H. Li et al., 2009).

Mutation Verification

1. Primer3 software (distributed under GPLv2) for primer design and primers (Rozen & Skaletsky, 2000).
2. Plate Spinner Centrifuge.
3. Commercial 10mM Tris-HCl, pH=8.5 buffer.
4. 96-well PCR plates.
5. Polymerase Chain Reaction (PCR) materials: QIAGEN® Taq PCR Master Mix Kit or QIAGEN® LongRange PCR Kit and Strain Genomic DNA (from material 2.1.1).
6. Gel loading materials: Blue (6X) Gel Loading Dye and ethidium bromide, 1% Solution/Molecular Biology, for visualization of PCR products and 1 Kb Plus DNA Ladder for size determination.
7. 50X TAE: 242g Tris base, 57.1ml Glacial Acetic Acid, 18.6g EDTA dissolved in 900mL nanopure water. Add make up nanopure water to 1L.
8. TAE DNA gel for separating DNA fragments: dissolve 1% W/V Agarose in 1X TAE.
9. Gel electrophoresis equipment.

10. PCR Purification Kit to purify PCR products.
11. DNA sequence finishing software Phred/Phrap/Consed or CodonCode Aligner (Ewing & Green, 1998; Ewing, Hillier, Wendl, & Green, 1998; D. Gordon et al., 1998).
12. Thermal Cycler for generating PCR products.

Methods

Obtaining the evolved strain and interpretation of mutation function is outside the scope of this paper. Here we restrict this protocol to DNA purification, genome sequencing, analysis and verification.

Obtain Sequence Data

1. After obtaining an evolved bacteria colony isolate, prepare to use the QIAGEN DNeasy® blood & tissue kit. Other commercial kits can also be used to isolate the genomic DNA. First, grow the parent strain (before the evolution experiment) and the evolved strain overnight in 25 mL LB.
2. Follow the QIAGEN DNeasy® Blood & Tissue kit protocol for gram-negative bacteria.

    2.1. Harvest cells (maximum 2 x 109 cells) in 50mL centrifuge tube by centrifuging for 20 minutes at 4°C, ~5,000xg. Discard supernatant (see Note 1).

    2.2. Resuspend pellet in 180 μl Buffer ATL and transfer to a microcentrifuge tube.

    2.3. Add 20 μl proteinase K. Mix thoroughly by vortexing, and incubate at 56°C in a temperature controlled waterbath until the cells are completely lysed (3h). Vortex *every hour*.

2.4. Add 20 μl RNase A, briefly vortex, and incubate at room temperature for 2 minutes.

2.5. Add 200 μl Buffer AL, and mix thoroughly by vortexing. Then add 200 μl 100 % ethanol and mix again thoroughly by vortexing (see Note 2).

2.6. Pipet the sample into the DNeasy Mini spin column and centrifuge at maximum speed for 1 minute. Discard flow-through (see Note 3).

2.7. Add 500 μl Buffer AW1 and centrifuge at maximum speed for 1 minute.

2.8. Add 500 μl Buffer AW2, and centrifuge at maximum speed for 1 minute. Discard flow-through and centrifuge again at maximum speed for 1 minute to dry the column (see Note 4).

2.9. Place the DNeasy Mini spin column in a clean 1.5 ml microfuge tube, and pipet 100 μl Buffer AE directly onto the DNeasy membrane. Incubate at room temperature for 1 min, and then centrifuge at maximum speed for 1 min to elute. Add another 100 μl Buffer AE, incubate for 1 minute, and then centrifuge at maximum speed for 1 minute (see Note 5). Freeze DNA at -20°C or proceed directly to the next step.

3. Check the quality of the genomic DNA on a nanodrop. First, blank the spectrophotometer with 1 μl nanopure water. Wipe away the water, then add the sample. One should see a smooth profile with a minimum at 230nm and a maximum at 260nm. Typical values should be ~20 μg genomic DNA, 280/260 value of ≥ 1.8, and a 260/230 value of ≥ 2. If the quality is too low, repeat the wash steps 2.7-2.9 with a new column.

4. Submit ≥ 2 μg/sample genomic DNA (at least 1 parent strain sample and 1 evolved strain sample) to a core facility for whole genome sequencing. There are many options to choose which sequencing instrument and which sequencing method; currently the DNA

core facility at Iowa State University has a GAII sequencer from Illumina, INC. 75-cycle paired-end sequencing is recommended as the researcher obtains more reads at a higher quality. To date, Illumina offers 150-cycle paired-end data with the GAII sequencer and 100-cycle paired-end data on their HiSeq instrument. If submitting more than one sample, indexing is the best option as one pays only for one sequencing lane. Indexing allows to a maximum of 12 samples in a single lane. The workflow of the Illumina platform is shown in Figure 1. Refer to the Illumina website for their sequencing technology: http://www.illumina.com/technology/sequencing_technology.ilmn (see Note 6).

Preprocess Sequence Data

High throughput sequence data is most commonly stored in FASTQ format. FASTQ format represents each read as a set of lines: header, sequence, sequence ID (optional), and quality scores in ASCII encoding. These text files typically have a .fq, .fastq, or .txt extension.

1. Before beginning analysis, identify what quality scoring encoding is associated with the raw data. Different Illumina genome analyzer pipeline software versions use different scoring scheme variations (e.g., Illumina 1.3+, Illumina 1.5+, and Illumina 1.8+). If there is difficulty identifying which encoding is used, FastQC includes this in its output. Software user manuals will specify if a particular scoring scheme is expected as input and it may be necessary to perform a conversion prior to analysis using Galaxy, the FASTX-Toolkit, or using a “Bio*” library in the language of your choice (e.g., BioPerl, BioPython, BioRuby, BioJava).

2. Use FastQC to perform an initial quality assessment of raw data. Launch the FastQC GUI and open FASTQ data files to generate all FastQC reports at once. Reports and graphs are presented in HTML format and can be saved for reference.

   Examine per-base quality, per-sequence quality, per-base content, and length distributions (not applicable for Illumina reads). Also check for overrepresentation of sequences and if they correspond to contaminants or PCR artifacts (in addition to common artifacts provided by FastQC users may supply sequences of potential contaminants to screen for).

   Use the summary icons (green: normal, orange: slightly abnormal, and red: very unusual) as guidelines in the following preprocessing steps. It is important to acknowledge that not all preprocessing steps will be necessary for all data and also that having small abnormalities may be acceptable in the context of the data and should not prevent a researcher from proceeding with analysis.

3. Perform read trimming using the FASTX-Toolkit if necessary. Read quality deteriorates with position and base calls near the end of a read are more prone to error. An appropriate length to trim may be determined from FastQC output. Use the `fastx-trimmer` command from the FASTX-Toolkit:

   ```
   $ fastx_trimmer [-f N] [-l N] [-i INFILE] [-o
   OUTFILE]
   ```

   Where `[-f N]` specifies the first base to keep (default is 1), `[-l N]` specifies the last base to keep (default is entire read), `INFILE` specifies the FASTQ file and `OUTFILE` is the name to give the trimmed data file. More advanced techniques allow for adaptive read trimming, however reads of varied length may not be acceptable as input for all analysis software.

4. Filter reads by overall quality with the FASTX-Toolkit:

   ```
   $ fastq_quality_filter [-q N] [-p N] [-i INFILE]
   [-o OUTFILE]
   ```

   The minimum quality score to keep is `[-q N]` and `[-p N]` is the minimum percentage of bases that must have `[-q]` quality.

5. Remove sequencing artifacts, described as reads that are predominantly one base (e.g., AAAAAAAAAAAAAAAACAAACA), using the FASTX-Toolkit:

   ```
   $ fastx_artifacts_filter [-i INFILE] [-o OUTFILE]
   ```

   Where `INFILE` specifies the FASTQ file and `OUTFILE` is the name to give the filtered data file.

6. Remove adapters sequences (identified as overrepresented sequences in the FastQC report or defined in protocol) with the FASTX-Toolkit:

   ```
   $ fastx_clipper [-a ADAPTER] [-l N] [-i INFILE] [-
   o OUTFILE]
   ```

   Where `[-a ADAPTER]` is the adapter sequence that is to be removed from 3'-end of sequences, `[-l N]` is the minimum length of reads to keep in the dataset (default is 5), `INFILE` specifies the FASTQ file, and `OUTFILE` is the name to give the filtered data file.

7. Resubmit filtered and trimmed data to FastQC to verify improved data quality and recalculate data summary statistics before proceeding with analysis.

Map Short Reads to Reference Genome

1. Build a reference index (using the reference genome in FASTA format) using the alignment algorithm of your choice, e.g.,

   ```
   $ bwa index [-p prefix] [-a algoType] ref.fa # BWA
   $ bowtie-build [options]* ref.fa <prefix> # Bowtie
   $ bowtie2-build [options]* ref.fa <prefix> #Bowtie 2
   ```

   Where `ref.fa` is the reference genome in FASTA format, prefix is the prefix of the output database and also the database filename. Additional options are defined in the corresponding user manuals. Using the genome of the parent strain as the reference yields the best alignments. If the genome of the parent strain has not been sequenced, download the genome of the wild-type laboratory strain from a public online database such as NCBI RefSeq. One benefit of using the wildtype genome as reference is the ability to easily leverage existing annotation in publically available databases (e.g., BioCyc).
2. Align reads to the reference and generate a SAM file. The SAM file format is a TAB-delimited text file that contains information such as alignment position (or '*' for unaligned reads) and mapping quality for each read and is the common output format for aligners. SAMtools performs conversions between SAM and a compressed and indexed binary format called BAM.
3. Assess overall alignment quality by reviewing the summary statistics generated by mapping software such as the percentage of reads that aligned to the reference genome. Use SAMtools to calculate read depths for each position of the genome:

   ```
   $ samtools depth aln.sorted.bam > depth.txt
   ```

   Where `aln.sorted.bam` is the sorted BAM file. The output file, `depth.txt`, contains one line for each position in the reference genome. The second column is the coordinate and the third column is the number of reads that cover that position. The SAMtools depth utility does not report positions where read depth is zero, thus the number of lines in the file is equal to the number of bases where coverage is non-zero. Alternative, specify a depth cutoff to ignore very small read depths (i.e., do not consider depth = 1 as genome coverage). Calculate base coverage with one of the following:

```
$ wc -l depth.txt # depth > 0
$ awk '$3 > $N {i++} END{print i}' depth.txt# depth > N
```

Divide base coverage by genome size to obtain the percentage of the genome covered by reads.

Map quality scores can also be examined by investigating column 5 (MAPQ) of the SAM file.

***De novo* Assembly**

De novo assembly and mapping of short reads to a reference sequence are fundamentally different analysis procedures. Assembly of the genomes of evolved bacterial strains can be used to search for novel insertions and complex mutations that are difficult for mapping software to identify. Additionally, results from assembly methods can provide support for proposed alignments.

Assembly of short read data does not use a reference sequence and instead tiles reads to generate sequences called contigs. Incorporating the average distance between paired-end reads (called the insert size) is used to join contigs into scaffolds. The most important parameter in de Bruijn graph based assembly algorithms is the hash length, which is also known as the *k*-mer length. Large *k*-mer values require longer overlap between reads in order for them to be assembled (therefore the *k*-mer value may not be larger than the read length).

Conversely, small *k*-mer values require short overlap which results in increased sensitivity, but decreased specificity. The experimenter must provide the *k*-value parameter and there is no method to find the optimal value. Because of this, it is recommended that researchers test multiple *k*-mers and then compare several assemblies before proceeding.

1. Assemble short read data with the assembler of your choice. Test multiple *k*-mer values and calculate the total number of contigs, N50, and N90 for each assembly (typically reported by assembly software).
2. Proceed with the "best" assemblies such that the number of contigs is minimized and the N50 and N90 are maximized.

Identify Variations in Evolved Strains

1. Identify single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) for an alignment using SAMtools/BCFtools:

   ```
   $ samtools mpileup -uf ref.fa aln.bam | bcftools view -bvcg
   - > var.raw.bcf
   ```

   Where `ref.fa` specifies the reference genome in FASTA format and `aln.bam` is a binary alignment file. The output is a binary file (BCF) for Variant Call Format (VCF) TAB-delimited files. VCF is standard for storing information about variants in alignment data.

2. Find large deletions by inspecting of read depth (the number of reads mapped to a specific position on the reference genome). Calculate read depth values using SAMtools:

   ```
   $ samtools idxstats aln.bam
   ```

   Regions with zero or very low read depth may indicate deletions. Determining what qualifies as "low" read depth may be aided by examining the read depth distribution.

3. Use assembly results to distinguish complex mutations such as large insertions, duplications, and inversions that are difficult for mapping algorithms to identify. Align contigs to a reference genome or an alignment consensus sequence. First, generate the consensus sequence from an alignment file with SAMtools:

   ```
   $ samtools mpileup -uf ref.fa aln.bam | bcftools view -cg -
   | vcfutils.pl vcf2fq > cns.fq
   ```

   Where `cns.fq` is the output consensus sequence. Next, BLAST contigs against these sequences to reveal sequence variations. Syntenic dotplots can also be used to visually identify discontinuities.
4. If possible, leverage reference genome annotation to form hypotheses about the effects of mutation. Verification by targeted sequencing can be used to confirm mutations. More advanced experimentation is necessary to confirm hypothesized effects.

Verify Mutations

1. Obtain a list of mutations from the above analysis and the sequences of the regions of interest.
2. There are two approaches for obtaining primers for PCR: for genes and for noncoding regions.

   2.1. For mutated regions containing open reading frames (ORFs or genes), first note how large the gene is and round up to the nearest 1,000. Add the additional sequences upstream and downstream of the gene equally. Then split the sequence into 1,000 bp segments (see Note 7 and Note 8). This method will give you room to pick optimal primers to include the entire sequence of interest. Use the Primer3 program to design optimal primers whose PCR product size range is 851-1,000 bp in length (see Note

9). Paste in the first 1,000 bp sequence, use the other default values, and click "Pick Primers". Select any of the suggested primers, noting where they bind to the template and the product size. Add the next 1,000 bp block of sequence and repeat until complete.

2.2. For mutated regions within a non-coding region (NCR), take a 1,000bp segment of DNA sequence and set the suspected mutation in the middle (~500bp from the first base). This ensures good sequencing data of this region. Use the Primer3 program for NCRs the same way for ORFs (see Note 10).

3. For long sequencing regions (>1kb), the above method will have gaps in the total sequence. In order to fill in the gaps, repeat the process with a 500bp offset and choose the reverse complement primers that bind in the middle of the sequence. This will also increase the fidelity of the sequence data (see Figure 2).
4. After choosing the primers, order them from an oligo synthesis company. If you have multiple primers, a 96-well plate format may be convenient. Resuspend them in either nanopure water or 10mM Tris-HCl, pH=8.5 at 100μM, vortex, and centrifuge briefly.
5. Keep all PCR materials on ice and set up your PCR reaction in a 96-well plate according to Table 1. Reserve one well for a negative control PCR (no template, choose any primer pair) to check for contamination. Run PCR using a thermocycler according to Table 2. If the sequencing region is longer than 1kb, it is possible to make a long PCR product and then submit multiple sequencing primers for a single template (up to 5kb) (see Note 11). For higher fidelity, especially at longer sequencing templates (>5kb) or difficult templates (high GC content), use the QIAGEN® LongRange PCR kit according to Table 3 and Table 4.

6. Check the concentration of the PCR products using a nanodrop (see 3.1.3). Check the size of the PCR product on a 1% TAE agarose gel.

    6.1. Melt 1X TAE with 1% agarose gel in a standard microwave. Add 25mL with 2 drops of ethidium bromide to a 8.5x10 cm gel casting tray in a gel casting tray holder with either an 8 or 15 sample comb (see Note 12). For more samples, use a 17x10 gel casting tray with a 26 sample comb. In this case, use 50mL of 1% TAE agarose gel with a few drops of ethidium bromide. Allow 30 minutes for the gel to solidify.

    6.2. Remove the comb and the gel casting tray. Place the gel casting tray into the gel box. Add 1X TAE until the surface of the gel is covered evenly.

    6.3. Mix standard and samples according to Table 5. Mix the standard and load into the first well; perform the same with each sample. Set the voltage to 100, put the top on, and click “run” (see Note 13). Wait 45 minutes - 1 hour for the dye to reach the bottom. Turn the system off when finished.

    6.4. Use the UV camera to visualize the PCR products. Match the standard with the PCR product to determine the approximate size (see Note 14).

7. If the PCR worked as expected, submit samples for sequencing by a core facility or company. The sample may need to be purified (use a standard PCR Purification Kit protocol) before submission (check the submission requirements). Use the sequencing primers as described in Figure 2 (see Note 15). The sequencing data is returned as .ab1 trace files and .seq files. One can view the .seq files in any text editor program. More advanced analysis requires the use of .ab1 trace files and DNA sequence finishing software.

8. In CodonCode Aligner (or any sequence finishing software), load the forward and reverse sequence of the samples. The first 20 bases and the last few bases (depends on the sequence length) have low quality scores. Highlight the samples and choose "clip ends" using the default parameters. Highlight the overlapping samples and assemble them into contigs (see Figure 3). The consensus sequence is shown at the bottom, where the base with the highest quality score is chosen. Here one can manually edit the sequence and call individual bases that are difficult. If there are discrepancies, open the trace files again to determine which is correct (see Note 16).
9. Use the BLASTn alignment tool (choose "Align two or more sequences" option) to align the consensus sequence to the parent strain and wildtype sequence. Sequences that are not matching or are unknown can be found using the NCBI nucleotide BLAST database.

Notes

1. The methods described here are developed in our lab, unless it is a published protocol from Illumina, INC. or commercial kit protocols. The steps using commercial kits are the published protocols of the kit manufacturer, where special deviations are in italics. Harvesting cells at 4°C, 4,000rpm prevents lysis and increases DNA yields. Do not overload the DNA column. Overloading the column causes blockage of the membrane and decrease yields. To obtain the maximum cell count per DNA column, first obtain a correlation of OD (we use 550 nm for *E. coli*) to *C* cells/mL (outside the scope of this protocol). Next, use to calculate the amount of cells you need:

$$\frac{2\ x\ 10^{9} cells}{c\ {}^{cells}/_{mL} x\ 10\ mL} = X\ OD_{550} \qquad (1)$$

For example, if C = 1.69 x 10^8, we have $\frac{2\ x\ 10^{9} cells}{1.69\ {}^{cells}/_{mL}\ x\ 10\ mL} = 1.18\ OD_{550}$. Therefore, 10 mL, OD_{550} 1.18 is required to obtain 2 x 10^9 cells.

2. The ethanol, sample and Buffer AL need to be mixed immediately and thoroughly by vortexing. Otherwise local precipitation may occur in the sample, which will decrease yields.
3. Buffer AL and Buffer AW1 are not compatible with bleach and may form decomposition products.
4. The column must be dried before eluting the DNA. Residual ethanol will decrease yields.
5. Subsequent elution steps will increase DNA yields, but decrease concentration. Do not elute more than 200 µl into a single 1.5 ml microfuge tube.
6. The insert sizes are less than 800 bp. We typically use 400-500 bp.
7. Due to possible polar effects from upstream mutations, the researcher may want to include the complete sequence from the promoter to the stop codon of the gene of interest. The sequencing length may be prohibitive and costly, so it is up to the researcher to include the upstream sequences along with the gene of interest. This is especially true if there are many genes in between the annotated promoter sequence and the gene of interest.
8. For example, the *E. coli* gene *carB* is 3,222 bp; therefore, round up to 4,000 bp by using this formula: 4,000-3,222=778/2=389 bp. Add 389 bp upstream and downstream of the *carB* gene. The total sequence is therefore 4,000 bp with the *carB* gene in the middle. This is enough to include a sequencing primer region and the promoter sequence 42bp from the translational start.
9. The limit of good quality Sanger sequence reads is about 1,000 bp. The Primer3 program chooses optimal primers and performs in silico PCR to obtain PCR products in the desired range. This ensures that each primer has approximately the same length and

melting temperature. It is good to also select alternate primers in case the primers weakly bind to the template. If the region is heavily mutated, it may be difficult choosing the correct primers.

10. NCRs may include long blocks of A-T rich sequences and therefore optimal primers may not be available. In this case, adjust the target sequence, so that the mutated region is closer to either the 5’ end or 3’ end. This way one can obtain optimal primers that can be used for sequencing this region.
11. QIAGEN® Taq DNA Polymerase is for general applications. For longer PCR products, the probability for incorporation of the incorrect base increases (false SNP); therefore, the use of a higher fidelity enzyme (i.e., QIAGEN® LongRange PCR kit) is recommended. Higher fidelity PCR enzymes are recommended for SNP identification and resequencing applications. It depends on the researcher which option is best. For extremely long PCR (10kb-40kb), the researcher is referred to the QIAGEN® LongRange PCR Handbook for an alternate PCR protocol.
12. Ethidium bromide is toxic and mutagenic. Always wear proper protection equipment.
13. Make sure that the diode colors match (black with black and red with red) and that the black one is at the top. This ensures that the DNA samples will run through the gel in the correct direction. Also the dye should not run off the gel, otherwise one may lose the samples.
14. If DNA bands are not visible, soak the gel for 1 hour in 1X TAE with ethidium bromide.

15. Use the following formula to calculate the number of sequencing reactions:

$$\#\text{Quality Sequences} = \frac{bp}{1000} * 2 - 1 \qquad (2)$$

16. Common mismatches occur when the local sequence contains blocks of the same base (i.e., A block of 6 A's in a row), or the ends are overlapping with one sample containing poor quality bases. This step may not be necessary as the consensus sequence is called according to quality. If SNPs or indels are discovered, this step becomes much more difficult, especially if there are duplication events.

Acknowledgements

We would like to thank Michael Baker at the DNA facility of the Iowa State University Office of Biotechnology for his input on next generation sequencing using the Illumina platform. We also thank Emily Rickenbach, an undergraduate student who helped automate the method for picking primers for verifying mutations. Funding was provided for this work by the NSF Engineering Research Center for Biorenewable Chemicals (CBiRC), NSF award number EEC-0813570 and NSF Energy for Sustainability award number CBET-1133319.

Refereences

Andrews, S. (2012). FASTQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/: Computer program distributed by author.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., . . . Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol, Chapter 19, Unit 19.10.11-21. doi:10.1002/0471142727.mb1910s89

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10, 421. doi:10.1186/1471-2105-10-421

Duro, A., & Serrano, R. (1981). Inhibition of succinate production during yeast fermentation by deenergization of the plasma membrane. Current Microbiology, 6(2), 111-113.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, 8(3), 186-194.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, 8(3), 175-185.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., . . . Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res, 15(10), 1451-1455. doi:10.1101/gr.4086505

Goecks, J., Nekrutenko, A., Taylor, J., & Team, G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11(8), R86. doi:10.1186/gb-2010-11-8-r86

Gordon, A. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html: Computer program distributed by the author.

Gordon, D., Abajian, C., & Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res, 8(3), 195-202.

Ingram, L. O., & Buttke, T. M. (1984). Effects of alcohols on micro-organisms. Adv Microb Physiol, 25, 253-300.

Jarboe, L. R. (2011). YqhD: a broad-substrate range aldehyde reductase with various applications in production of biorenewable fuels and chemicals. Appl Microbiol Biotechnol, 89(2), 249-257. doi:10.1007/s00253-010-2912-9

Jarboe, L. R., Grabar, T. B., Yomano, L. P., Shanmugan, K. T., & Ingram, L. O. (2007). Development of ethanologenic bacteria. Adv Biochem Eng Biotechnol, 108, 237-261. doi:10.1007/10_2007_068

Jørgensen, H., Kristensen, J., & Felby, C. (2007). Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. Biofuels, Bioproducts Biorefining, 1, 119-134.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25. doi:papers3://publication/doi/10.1186/gb-2009-10-3-r25

Li, C., Qian, W., & Zhao, Z. (2008). Acid in ionic liquid: an efficient system for hydrolysis of lignocellulose. Green Chemistry, 10(2), 177-182.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Subgroup, G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nat Rev Genet, 11(1), 31-46. doi:10.1038/nrg2626

Miller, E. N., Jarboe, L. R., Turner, P. C., Pharkya, P., Yomano, L. P., York, S. W., . . . Ingram, L. O. (2009). Furfural inhibits growth by limiting sulfur assimilation in ethanologenic Escherichia coli strain LY180. Appl Environ Microbiol, 75(19), 6132-6141. doi:10.1128/AEM.01187-09

Miller, E. N., Jarboe, L. R., Yomano, L. P., York, S. W., Shanmugam, K. T., & Ingram, L. O. (2009). Silencing of NADPH-dependent oxidoreductase genes (yqhD and dkgA) in furfural-resistant ethanologenic Escherichia coli. Applied and Environmental Microbiology, 75(13), 4315-4323. doi:papers3://publication/doi/10.1128/AEM.00567-09

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol, 132, 365-386.

Schwarz, K. M., Kuit, W., Grimmler, C., Ehrenreich, A., & Kengen, S. W. (2012). A transcriptional study of acidogenic chemostat cells of Clostridium acetobutylicum--cellular behavior in adaptation to n-butanol. J Biotechnol, 161(3), 366-377. doi:10.1016/j.jbiotec.2012.03.018

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res, 19(6), 1117-1123. doi:10.1101/gr.089532.108

Smith, E. H., Janknecht, R., & Maher, L. J. (2007). Succinate inhibition of alpha-ketoglutarate-dependent enzymes in a yeast model of paraganglioma. Hum Mol Genet, 16(24), 3136-3148. doi:10.1093/hmg/ddm275

Trinh, C. T., Huffer, S., Clark, M. E., Blanch, H. W., & Clark, D. S. (2010). Elucidating mechanisms of solvent toxicity in ethanologenic Escherichia coli. Biotechnol Bioeng, 106(5), 721-730. doi:10.1002/bit.22743

Winkler, J., & Kao, K. C. (2011). Transcriptional analysis of Lactobacillus brevis to N-butanol and ferulic acid stress responses. PLoS One, 6(8), e21438. doi:10.1371/journal.pone.0021438

Zaldivar, J., & Ingram, L. O. (1999). Effect of organic acids on the growth and fermentation of ethanologenic Escherichia coli LY01. Biotechnol Bioeng, 66(4), 203-210.

Zaldivar, J., Martinez, A., & Ingram, L. O. (1999). Effect of selected aldehydes on the growth and fermentation of ethanologenic Escherichia coli. Biotechnol Bioeng, 65(1), 24-33.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 18(5), 821-829. doi:10.1101/gr.074492.107

Tables

Table 1. A Typical 20 µL PCR Reaction

Material	Stock Concentration	Amount to add	Final concentration
Nuclease-free water	-	Variable	-
Primer A	100 µM	0.1 µL	500 nM
Primer B	100 µM	0.1 µL	500 nM
Template	Variable	Variable	50-500 ng
2X QIAGEN® Taq PCR Master Mix	2X	10 µL	1X or 2.5U Taq + 200 µM dNTP

Table 2. Typical PCR Reaction Conditions

Step	Time	Temperature	Comments
1. Denaturation	4 min	94°C	Denaturation of template and primer-dimers
2. Denaturation cycle	0.5 min	94°C	Or 5°C below the lowest primer melting temperature
3. Annealing cycle	0.5 min	55°C	
4. Extension cycle	1 min/kb	72°C	
5. Repeat steps 2-4			Repeat 30 times
6. Final extension	10 min	72°C	
7. End	infinite	5°C	

Table 3. 20 µL QIAGEN® LongRange PCR Kit (up to 10kb) Setup

Material	Stock Concentration	Amount to add	Final concentration
Nuclease-free water	-	Variable	-
Primer A	10 µM (diluted 10 fold)	0.8 µL	400 nM
Primer B	10 µM (diluted 10 fold)	0.8 µL	400 nM
Template	Variable (diluted 10 fold)	Variable	0.1-10 ng
dNTP mix	10 mM of each base	1 µL	500 µM of each base
QIAGEN® LongRange PCR buffer	10X	2 µL	1X or 2.5U Mg^{2+}
QIAGEN® LongRange PCR enzym Mix	100U (total enzyme mix)	0.16 µL	0.8U

Table 4. QIAGEN® LongRange PCR Reaction Conditions

Step	Time	Temperature	Comments
1. Denaturation	3 min	93°C	Denaturation of template and primer-dimers
2. Denaturation cycle	15 s	93°C	Or 5°C below the lowest primer melting temperature
3. Annealing cycle	0.5 min	62°C	
4. Extension cycle	1 min/kb	68°C	
5. Repeat steps 2-4			Repeat 35 times
6. End	infinite	4°C	

Table 5. Recipe for Mixing Standard and Samples

Standard		Samples	
Standard	1 μL	PCR product	8 μL
Dye	2 μL	Dye	2 μL
Nuclease-free water	7 μL	Nuclease-free water	-

Figures

Figure 1. *Illumina sample preparation Protocol*

Illumina sample preparation protocol, adapted from the Illumina guide Preparing Samples for Sequencing Genomic DNA. See the Illumina guidebooks for their detailed protocols.

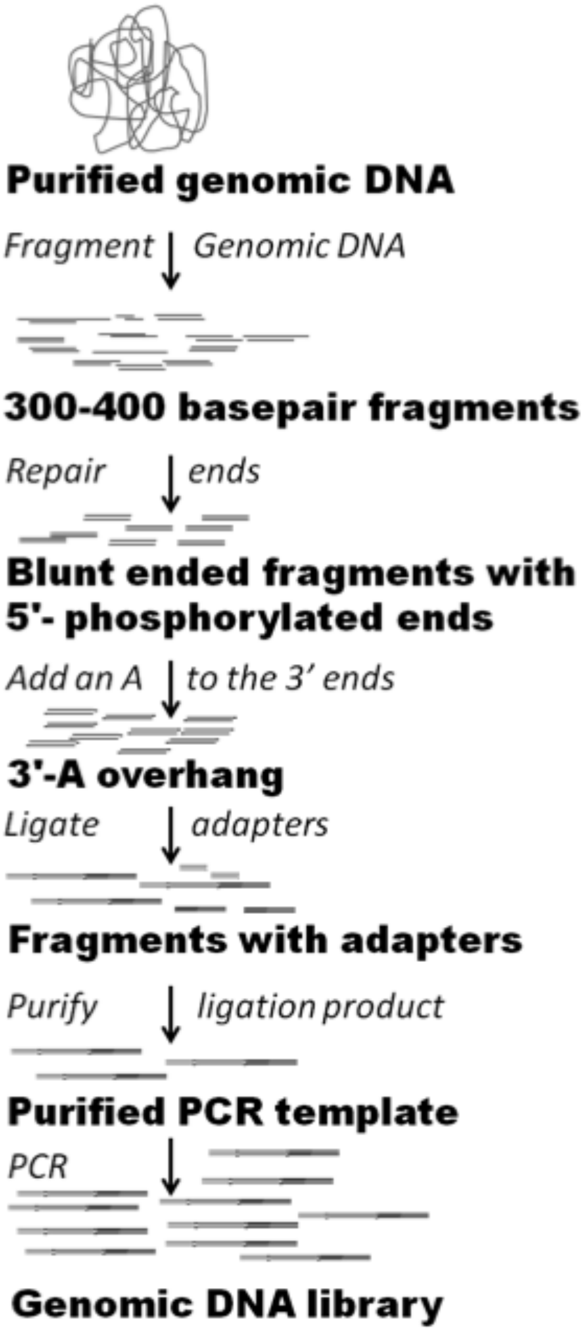


Figure 2. *Schematic for designing sequencing Primers for long sequencing regions*

The average quality score is plotted for each individual base along the template. A quality score of 20 or greater is considered acceptable. Choose the forward primer binding to the lagging strand to cover the general area (top). To fill the gaps, use the reverse primer binding to the leading strand, with a 500 bp offset. In this example, the forward sequencing primers will bind to bases 1, 1000, 2000, & 3,000. The reverse sequencing primer will bind to bases 3,500, 2,500, & 1,500.

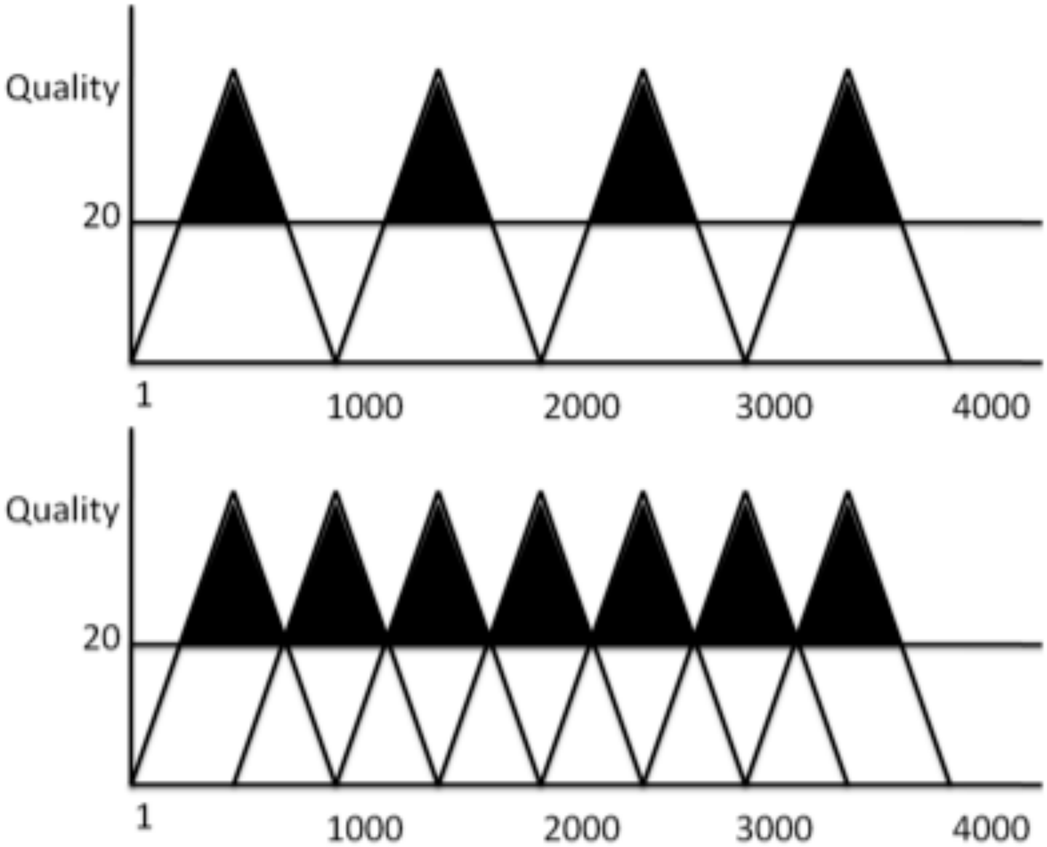
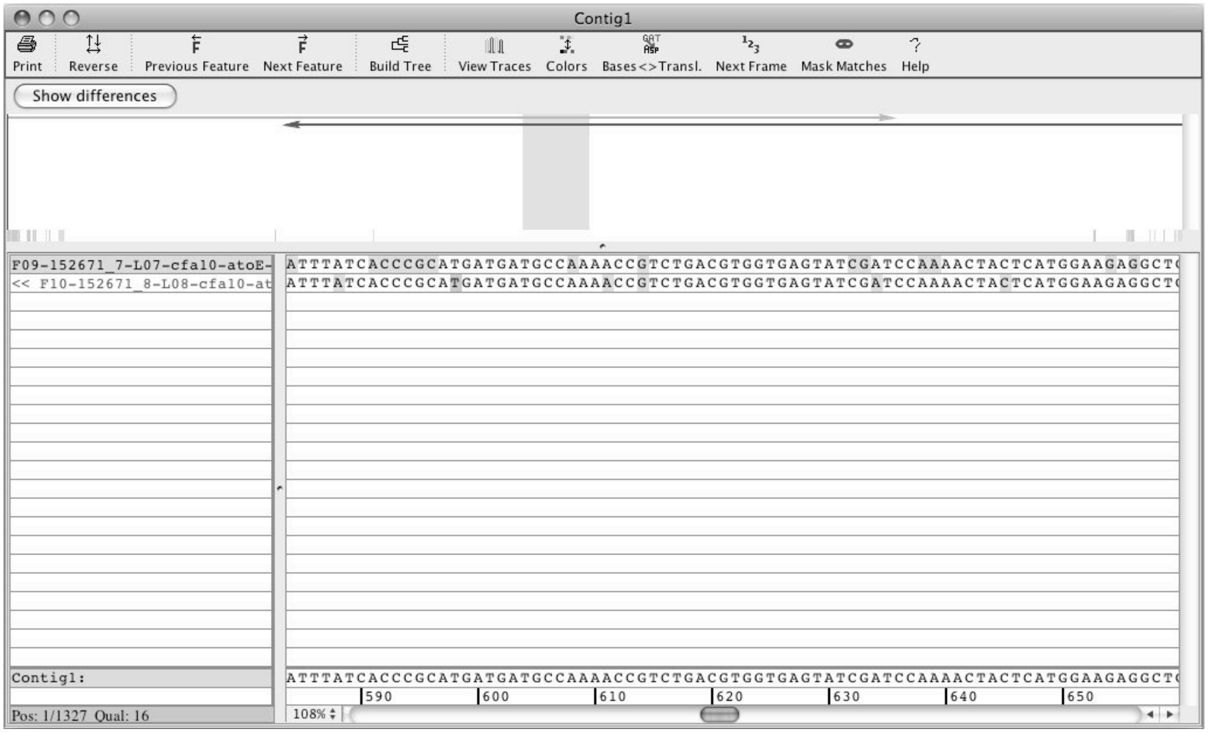


Figure 3. *Contig assembly in CodonCode Aligner*



```
F09-152671_7-L07-cfa10-atoE-  ATTTATCACCCGCATGATGATGCCAAAACCGTCTGACGTGGTGAGTATCGATCCAAAACTACTCATGGAAGAGGCT
<< F10-152671_8-L08-cfa10-at  ATTTATCACCCGCATGATGATGCCAAAACCGTCTGACGTGGTGAGTATCGATCCAAAACTACTCATGGAAGAGGCT

Contig1:                      ATTTATCACCCGCATGATGATGCCAAAACCGTCTGACGTGGTGAGTATCGATCCAAAACTACTCATGGAAGAGGCT
                                   |590      |600      |610      |620      |630      |640      |650
Pos: 1/1327  Qual: 16
```

APPENDIX B. *E. COLI* VARIANT ANALYSIS (EVA) USER GUIDE

Software Requirements

EVA requires Java to be installed as well as access to a copy of the RegulonDB database and the EVA supplementary database. Additionally, Provean and Vienna RNA packages must be installed for various analyses to be performed. Networks generated by EVA can be viewed using various visualization software. EVA provides a custom style for use with Cytoscape.

Commands

EVA: *E. coli* Variant Analysis 1.0

Annotate, analyze, and prioritize *E. coli* mutations in comma separated value (.csv) files, text (.txt) files, and GenomeDiff (.gd) files.

```
Usage: java -jar eva.jar -d derivative1.gd [-d derivative2.gd...] [-p parent1.gd...] [options]
       java -jar eva.jar -r N [options]
```

Options:

File

-d, --derivative <file_path>   path to CSV file containing sequence variations in derivative (i.e., evolved) strain

-p, --parent <file_path>   Path to CSV file containing sequence variations in parent (i.e., ancestral) strain

-r, --random <integer> Run EVA for N random mutations instead of importing from file(s)

-o, --output <path> Path to desired project output directory (DEFAULT = output)

Annotation

--exact Annotate all features in large mutations that regulate 1+ features outside the mutation (DEFAULT = ignore non-coding features in large mutations)

--large_mutation <integer> The size (in bp) for which to consider a mutation "large" (DEFAULT = 1000)

Analysis

--noRNA Do not predict optimal secondary structures or minimum free energy for RNA features

--noAA Do not run PROVEAN analysis on variant amino acid sequences

--provean_threshold <double> Cutoff PROVEAN score for assigning HIGH priority. Scores less than or equal to this value are predicted to be damaging. (DEFAULT = -2.5)

--num_threads <integer> Number of threads to use in BLAST search (DEFAULT = 1)

--maxRNA <integer> Maximum length of RNA to submit for RNA analysis. (DEFAULT = 1000)

--deltaMFE <double> Cutoff change in MFE for assigning HIGH priority, in standard deviations. Final cutoff is calculated per RNA sequence. Cannot be used with --edit_distance. (DEFAULT = 1)

--edit_distance <integer> Use Levenshtein distance between predicted optimal structures instead of free energy calculations. Cannot be used with --deltaMFE_threshold.

Network

--full_network Construct network with all features and links (DEFAULT = abbreviated network that consolidates select interactions)

--regulatory_steps <integer> Number of regulatory steps to incorporate when building network (DEFAULT = 1)

--derivative_networks In addition to a network formed from all samples, construct a network for each derivative from mutations that are not common to that derivative and parent sample(s)

Help

-h, --help Print help and exit

Multiple parent and derivative files may be submitted at once. At least one derivative file is required to run EVA.

CSV files must contain one mutation per line:

```
    POS,REF,ALT
```

where POS is the position on the genome, REF is the reference DNA sequence, and ALT is the alternate DNA sequence.

Output files for mutation analysis and network generation can be found in the specified output directory.

EVA.out is a tab delimited text file containing the results of the EVA analysis. Additional files for network and PROVEAN analysis are located in their respective directories.

Networks are formed from seed nodes representing features that correspond to mutations that are not present in all samples (if more than one sample is imported).

Additional configuration options can be found in the config.properties file.

Configuration

Config.properties file:

```
# The name of the installed RegulonDB database
RegulonDB_database=regulondb94

# The name of the installed EVA database (e.g., eva)
EVA_database=eva

# The server to connect to for database access (e.g., localhost)
server=localhost

# The port number
port=8889

# The username to access the databases
username=root
```

```
# The password to access the databases
password=root

# Absolute path to provean.sh from PROVEAN installation (e.g.,
/usr/local/bin/provean.sh)
PROVEAN=/usr/local/bin/provean.sh

# Path to compiled PROVEAN library (available with EVA in ProveanLibrary
directory)
PROVEAN_library=/path/to/ProveanLibrary

# Directory containing Vienna RNA executables (e.g., /usr/local/bin/)
ViennaRNA=/usr/local/bin/

# Output fields. * denotes fields automatically included in output
regardless of configuration.
# *   POS               Position in the reference genome
# *   REF               Reference sequence (DNA)
# *   ALT               Alternate sequence (DNA)
#     VARIATION         Type of sequence variation
#     SPAN              How the sequence variation corresponds to an
annotation
#     GENE_MUTATION     Type of gene mutation
# *   FEATURE_ID        Feature identifier in RegulonDB or EcoCyc database
# *   FEATURE_TYPE      Type of feature corresponding to the mutation
# *   FEATURE_NAME      Name of Feature
#     BNUM              B-number (only applicable for genes)
#     STRAND            Feature strand (+ corresponds to forward strand, -
corresponds to reverse strand)
#     FEATURE_LEFT      Leftmost position of feature
#     FEATURE_RIGHT     Rightmost position of feature
#     FEATURE_REF       Feature reference sequence (DNA)
#     FEATURE_ALT       Feature alternate sequence (DNA)
#     FEATURE_AA_REF    Feature reference sequence (AA, only applicable
for genes)
#     FEATURE_AA_ALT    Feature alternate sequence (AA, only applicable
for genes)
#     SEQ_EDIT_DIST     The Levenshtein distance between the reference and
alternate feature sequence
#     STRUCTURE_DIST    The edit distance between reference and alternate
feature predicted structures (only applicable for RNA features)
#     HGVS              Description of mutation following Human Genome
Variant Society nomenclature, used for PROVEAN analysis (only applicable
for protein-coding gene features)
#     PROVEAN           PROVEAN score (only applicable for protein-coding
gene features)
#     RNA_THRESHOLD     Delta MFE threshold for assigning high priority
#     RNA_MFE_REF       Predicted MFE for reference sequence (only
applicable for RNA features)
#     RNA_MFE_ALT       Predicted MFE for alternate sequence (only
applicable for RNA features)
#     RNA_STRUCT_REF    Predicted secondary structure for reference
sequence (only applicable for RNA features)
#     RNA_STRUCT_ALT    Predicted secondary structure for alternate
sequence (only applicable for RNA features)
#     SIGMA70_REF       Sigma70 predicted MFE for reference sequence (only
applicable for promoter features associated with sigma 70)
#     SIGMA70_ALT       Sigma70 predicted MFE for reference sequence (only
applicable for promoter features associated with sigma 70)
# *   PRIORITY          Priority assigned to mutation (UNASSIGNED, LOW,
HIGH)
# *   COMMENT           EVA comments and errors running analysis

# User-specified fields to include in output file.
fields=VARIATION,\
SPAN,\
GENE_MUTATION,\
BNUM,\
STRAND,\
FEATURE_LEFT,\
FEATURE_RIGHT,\
FEATURE_REF,\
FEATURE_ALT,\
FEATURE_AA_REF,\
FEATURE_AA_ALT,\
SEQ_EDIT_DIST,\
STRUCTURE_DIST,\
HGVS,\
PROVEAN,\
RNA_THRESHOLD,\
RNA_MFE_REF,\
RNA_MFE_ALT,\
RNA_STRUCT_REF,\
RNA_STRUCT_ALT,\
SIGMA70_REF,\
SIGMA70_ALT
```

Hierarchy for all packages

Class Hierarchies

- java.lang.Object
  - annotation.AbstractAnnotation<T> (implements java.lang.Comparable<T>)
    - annotation.impl.Annotation
    - annotation.impl.GeneAnnotation
    - annotation.impl.PromoterAnnotation
    - annotation.impl.RNAAnnotation
    - annotation.impl.Unannotated
  - dao.AbstractDatabaseDAO
    - dao.impl.EvaDBDAO
    - dao.impl.RegulonDBDAO
  - feature.AbstractFeature
    - feature.AbstractGenomicFeature
      - feature.AbstractPromoterFeature
        - feature.impl.Box10
        - feature.impl.Box35
      - feature.impl.AttenuatorTerminator
      - feature.impl.Gene
      - feature.impl.Operon
      - feature.impl.Rfam
      - feature.impl.ShineDalgarno
      - feature.impl.SRNAbs

    - feature.impl.Terminator
    - feature.impl.Tfbs
  - feature.impl.Attenuator
  - feature.impl.Pathway
  - feature.impl.Product
  - feature.impl.Reaction
  - feature.impl.TranscriptionFactor
  - feature.impl.TranscriptionUnit
- dao.AbstractFeatureDAO
  - dao.impl.AttenuatorDAO (implements dao.FeatureDAO)
  - dao.impl.AttenuatorTerminatorDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.Box10DAO (implements dao.GenomicFeatureDAO)
  - dao.impl.Box35DAO (implements dao.GenomicFeatureDAO)
  - dao.impl.GeneDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.OperonDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.PathwayDAO (implements dao.FeatureDAO)
  - dao.impl.ProductDAO (implements dao.FeatureDAO)
  - dao.impl.ReactionDAO (implements dao.FeatureDAO)
  - dao.impl.RfamDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.ShineDalgarnoDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.SRNAbsDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.TerminatorDAO (implements dao.GenomicFeatureDAO)
  - dao.impl.TfbsDAO (implements dao.GenomicFeatureDAO)

  - dao.impl.TranscriptionFactorDAO (implements dao.FeatureDAO)
  - dao.impl.TranscriptionUnitDAO (implements dao.FeatureDAO)
- network.AbstractNetwork
- result.AbstractResult
  - result.GeneResult
  - result.GenericResult
  - result.PromoterResult
  - result.RNAResult
  - result.UnannotatedResult
- analysis.AnalysisStrategyFactory
- annotation.AnnotationFactory
- analysis.BindingSiteStrategy (implements analysis.AnalysisStrategy<T>)
- sequence.Codon
- core.Consts
- network.Edge
- core.EVA
- util.fileImport.FastaReader
- dao.FeatureDAOFactory
- feature.FeatureLink
- java.util.logging.Formatter
  - util.ConsoleFormatter
- analysis.GenericStrategy (implements analysis.AnalysisStrategy<T>)
- analysis.GeneStrategy (implements analysis.AnalysisStrategy<T>)

- annotation.HGVS
- feature.Interaction
- genome.Interval
- genome.IntervalComparator (implements java.util.Comparator<T>)
- jdbc.impl.JdbcManagerImpl (implements jdbc.JdbcManager)
- util.Kmer
- util.LevenshteinAlignment
- genome.MergeIntervals
- mutation.Mutation (implements java.lang.Comparable<T>)
- mutation.MutationComparator (implements java.util.Comparator<T>)
- util.fileImport.MutationImporter
- network.Network
    - network.Branch
- network.NetworkBuilder
- network.Node
    - network.RootNode
        - util.fileImport.ParseCSV
- util.fileImport.ParseGD
- util.fileImport.ParseProvean
- util.fileImport.ParseVCF
- core.Project
- analysis.PromoterStrategy (implements analysis.AnalysisStrategy<T>)
- util.scripts.ProveanResult

- util.RandomMutation
- feature.Regulation
- util.ResultFormat
- util.scripts.RNAfoldResult
- analysis.RNAStrategy (implements analysis.AnalysisStrategy<T>)
- core.RunEVA
- util.scripts.RunProvean
- util.scripts.RunVienna
- sequence.Sequence (implements java.lang.Comparable<T>)
  - sequence.AA
  - sequence.DNA
  - sequence.RNA
- network.ShortestPath
- util.Statistics
- annotation.TestHGVS
- java.lang.Throwable (implements java.io.Serializable)
  - java.lang.Exception
    - analysis.AnalysisException
    - core.InitializationException
    - util.fileImport.MutationImportException
    - java.lang.RuntimeException
      - annotation.AnnotationException
      - dao.exception.DataAccessException

          - dao.exception.DataAccessConnectionFailureException
          - dao.exception.DataAccessSQLException
        - feature.FeatureException
        - mutation.MutationException
        - sequence.SequenceException
      - sequence.TranslationException
  - util.VariantGenerator

Interface Hierarchy

- analysis.AnalysisStrategy<T>
- dao.FeatureDAO
  - dao.GenomicFeatureDAO
- jdbc.JdbcManager
- jdbc.RowMapper<T>

Enum Hierarchy

- java.lang.Object
  - java.lang.Enum<E> (implements java.lang.Comparable<T>, java.io.Serializable)
  - util.ResultFormat.Columns
  - feature.impl.Terminator.TerminatorClass
  - feature.impl.Attenuator.AttenuatorType
  - feature.impl.Gene.GeneType
  - feature.impl.AttenuatorTerminator.AttenuatorTerminatorType
  - feature.impl.Product.ProductType

- feature.impl.Pathway.PathwayType
- core.Priority
- core.Consts.Database
- core.Consts.Genome
- core.Consts.RegulonDB
- sequence.Strand
- annotation.Span
- annotation.HGVS.HGVSType
- annotation.GeneMutationType
- annotation.TestHGVS.HGVSpattern
- network.Distance
- mutation.Mutation.Variation
- feature.Mechanism
- feature.Interaction.InteractionType
- feature.FeatureType
- feature.AbstractPromoterFeature.SigmaFactor