UNIVERSITY OF CALIFORNIA SAN DIEGO

Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data

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

in

Bioinformatics and Systems Biology

by

Patrick Victor Phaneuf

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Nathan E. Lewis, Co-Chair
Doctor Adam M. Feist
Professor Christian M. Metallo
Professor Justin R. Meyer
Professor Elizabeth A. Winzeler

2021

The dissertation of Patrick Victor Phaneuf is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am where I am and who I am in large part because of the guidance and generosity of others. I would like to take this opportunity to acknowledge the many who have helped me discover and pursue my passions, and have therefore contributed to the development of this dissertation.

I would like to express deep gratitude to Dr. Adam Feist for his mentorship and support. He has championed all of my academic and research endeavors as well as cared for my welfare since the beginning of my graduate school career. I would also like to express deep gratitude to Professor Bernhard Palsson for his mentorship and for providing me with opportunities to continue exploring my passions.

Aside from Dr. Feist and Professor Palsson, I have been fortunate to have the guidance and assistance of other exceptional professors, lab members, scientists, co-workers, and associates. At the Systems Biology Research Group and the Novo Nordisk Foundation Center for Biosustainability, I would like to thank James Yurkovich, Muyao Wu, Daniel Zielinski, Robin Cai, Anand Sastry, Zak King, Nathan Mih, Lawrence Yang, Justin Tan, Jean-Christophe Lachance, Colton Llyod, David Heckmann, Troy Sandberg, Josefin Johnsen, Connor Olson, Ify Aniefuna, Dennis Gosting, Dr. Sebastian Schulz, Chris Dalldorf, Dr. Amitesh Anand, Dr. Hyungyu Lim, Dr. Lei Yang, Dr. Henrique Machado, Dr. Se Hyeuk Kim, Emre Ozdemir, Elsayed Mohamed, Mohammad Radi, Tim Stapleton, Evelyn Travnik, Ying Hefner, Richard Szubin, and Marc Abrams. At the University of California San Diego, I would like to thank Professor Vineet Bafna, Professor Pavel Pevzner and members of my doctoral committee: Professor Nathan Lewis, Professor Christian Metallo, Professor Justin Meyer, and Professor Elizabeth Winzeler.

Previous to and early in my graduate studies, I had been fortunate to have mentors

who, among many things, revealed to me the far-reaching possibilities for those in pursuit of wonder. I'd like to thank Professor Bafna of UCSD, Professor Digh and Professor Zhao of Mercer University, and Keith Causey of the West Coast at large. I would also like to thank Dr. Greg Goddard, Dr. Greg Kaduchak, Glen Davis, Jace Akerlund, Michael Quick, Barry Hurt, and Sergiy Dets, who were and some still are of the Thermo Fisher Research and Development Group in Eugene, Oregon.

I would like to thank the institutions whose funding have supported this work. These include the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517), the National Institutes of Health's National Institute of Allergy and Infectious Diseases (U01AI124316), and the Institute for Systems Biology's Translational Research Fellows Program.

Last but not least, I would like to thank my family and friends for their love, support, and companionship.

Chapter 2 is a reprint of material published in: **Patrick V Phaneuf**, Patrick V Phaneuf, Dennis Gosting, Bernhard O Palsson, Adam M Feist, "ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation." Nucleic Acids Res. 2019 Jan 8;47(D1):D1164–71. The dissertation author is the primary author.

Chapter 3 is a reprint of material published in: **Patrick V. Phaneuf**, Patrick V. Phaneuf, James T. Yurkovich, David Heckmann, Muyao Wu, Troy E. Sandberg, Zachary A. King, Justin Tan, Bernhard O. Palsson, Adam M. Feist "Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity". BMC Genomics. 2020 Jul 25;21(1):514. The dissertation author is the primary author.

Chapter 4 is a reprint of material published in: **Patrick V. Phaneuf**, Daniel C. Zielinski, James T. Yurkovich, Josefin Johnsen, Richard Szubin, Lei Yang, Se HyeuK Kim, Christopher Dalldorf, Muyao Wu, Sebastian Schulz, Bernhard O. Palsson, Adam M. Feist "Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data." The dissertation author is the primary author.

VITA

2009	Bachelor of Science in Engineering, specialization: Computer, Mercer University
2016	Master of Science, Computer Science, University of California San Diego
2021	Doctorate of Philosophy, Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Mundhada, Hemanshu; Seoane, Jose M; Schneider, Konstantin; Koza, Anna; Christensen, Hanne B; Klein, Tobias; **Phaneuf, Patrick V**; Herrgard, Markus; Feist, Adam M; Nielsen, Alex T; 2017. Increased production of L-serine in Escherichia coli through adaptive laboratory evolution. *Metabolic engineering* 39:141-150.

Pekar, Jonathan E; **Phaneuf, Patrick**; Szubin, Richard; Palsson, Bernhard; Feist, Adam; Monk, Jonathan M; 2018. Gapless, unambiguous genome sequence for Escherichia coli C, a workhorse of industrial biology. *Microbiology resource announcements* 7(13).

Guzmán, Gabriela I; Olson, Connor A; Hefner, Ying; **Phaneuf, Patrick V**; Catoiu, Edward; Crepaldi, Lais B; Micas, Lucas Goldschmidt; Palsson, Bernhard O; Feist, Adam M; 2018. Reframing gene essentiality in terms of adaptive flexibility. *BMC systems biology* 12(1):1-20.

Phaneuf, Patrick V; Gosting, Dennis; Palsson, Bernhard O; Feist, Adam M; 2019. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. *Nucleic acids research* 47(D1):D1164-D1171.

Anand, Amitesh; Olson, Connor A; Yang, Laurence; Sastry, Anand V; Catoiu, Edward; Choudhary, Kumari Sonal; **Phaneuf, Patrick V**; Sandberg, Troy E; Xu, Sibei; Hefner, Ying; 2019. Pseudogene repair driven by selection pressure applied in experimental evolution. *Nature microbiology* 4(3):386-389.

Lloyd, Colton J; King, Zachary A; Sandberg, Troy E; Hefner, Ying; Olson, Connor A; **Phaneuf, Patrick V**; O'Brien, Edward J; Sanders, Jon G; Salido, Rodolfo A; Sanders, Karenina; 2019. The genetic basis for adaptation of model-designed syntrophic co-cultures. *PLoS computational biology* 15(3):e1006213.

Anand, Amitesh; Chen, Ke; Yang, Laurence; Sastry, Anand V; Olson, Connor A; Poudel, Saugat; Seif, Yara; Hefner, Ying; **Phaneuf, Patrick V**; Xu, Sibei; 2019. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. *Proceedings of the National Academy of Sciences* 116(50):25287-25292.

Du, Bin; Olson, Connor A; Sastry, Anand V; Fang, Xin; **Phaneuf, Patrick V**; Chen, Ke; Wu, Muyao; Szubin, Richard; Xu, Sibei; Gao, Ye; 2020. Adaptive laboratory evolution of Escherichia coli under acid stress. *Microbiology* 166(2):141-148.

Heckmann, David; Campeau, Anaamika; Lloyd, Colton J; **Phaneuf, Patrick V**; Hefner, Ying; Carrillo-Terrazas, Marvic; Feist, Adam M; Gonzalez, David J; Palsson, Bernhard O; 2020. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proceedings of the National Academy of Sciences* 117(37):23182-23190.

Phaneuf, Patrick V; Yurkovich, James T; Heckmann, David; Wu, Muyao; Sandberg, Troy E; King, Zachary A; Tan, Justin; Palsson, Bernhard O; Feist, Adam M; 2020. Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity. *BMC genomics* 21(1):1-16.

Sandberg, Troy E; Szubin, Richard; **Phaneuf, Patrick V**; Palsson, Bernhard O; 2020. Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. *Nature Ecology & Evolution* 4(10):1402-1409.

Lamoureux, Cameron R; Choudhary, Kumari Sonal; King, Zachary A; Sandberg, Troy E; Gao, Ye; Sastry, Anand V; **Phaneuf, Patrick V**; Choe, Donghui; Cho, Byung-Kwan; Palsson, Bernhard O; 2020. The Bitome: digitized genomic features reveal fundamental genome organization. *Nucleic Acids Research* 48(18):10157-10163.

Rychel, Kevin; Decker, Katherine; Sastry, Anand V; **Phaneuf, Patrick V**; Poudel, Saugat; Palsson, Bernhard O; 2021. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Research* 49(D1):D112-D120.

Phaneuf, Patrick V; Zielinski, Daniel C; Yurkovich, James T; Johnsen, Josefin; Szubin, Richard; Yang, Lei; Kim, Se HyeuK; Dalldorf, Christopher; Wu, Muyao; Schulz, Sebastian; Palsson, Bernhard O; Feist, Adam M; Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data. *In preparation.*

ABSTRACT OF THE DISSERTATION

Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data

by

Patrick Victor Phaneuf

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego 2021

Professor Bernhard Ø. Palsson, Chair
Professor Nathan E. Lewis, Co-Chair

Microbes are increasingly engineered for a large variety of valuable applications. Designing microbial strains remains challenging due to the complexity and knowledge gaps with biological systems. Bioengineers have a unique advantage over other engineering disciplines: biological systems have been solving challenges long before human intervention through adaptive evolution. Adaptive Laboratory Evolution (ALE) methods leverage this natural problem-solving process to elucidate biological functions and generate application solutions. The promise of ALE methods has resulted in the generation of a substantial amount of public ALE data, which in

aggregate, could contribute new insights towards ALE-derived strain design. The work of this dissertation combined aggregated public ALE data and rich mutation annotations to design strain variants with possible value in applications. Public ALE mutational data was consolidated into a web-accessible database that can report and export the aggregated ALE data. ALE metadata was used to statistically associate mutations to experimental conditions, reducing the amount of conditions to consider during mutation functional analysis and therefore deconvoluting the context of adaptive mutations. Multi-scale and multi-dimensional mutation annotations were used to identify the mutation effects, structural features, regulatory features, and cellular subsystems that ALE mutations converged upon. Meta-analysis of the aggregated ALE data revealed the principles underlying adaptive mutation trends, which were then used to design novel strain variants. The results of this dissertation demonstrate how strain design principles can be extracted from aggregated ALE data to enhance microbial engineering efforts.

Chapter 1

Introduction

The advent and continual improvement of biomolecular and bioinformatic technologies are fueling discovery and application in fields previously dominated by theory. High throughput sequencing has enabled an enormous breadth of organisms to be sequenced [1]. The development and application of DNA manipulating molecular machinery has enabled for DNA editing [2]. Further technologies and software algorithms have enabled the characterization of cellular systems and their components encoded on the genome, inviting more systematic approaches to manipulating cells through their genomes [3]. Multiple fields have emerged to take advantage of the newly acquired control over biology with different applications in mind. The field of synthetic biology has formed to encompass the efforts in implementing synthetic biological components, systems, and even organisms [4]. The field of metabolic engineering views biology as a manufacturing technology and works to control the titer, accumulation rate, and yield of specific metabolites production from microorganisms in an industrial setting [5]. The microbial strains generated through metabolic engineering are typically referred to as microbial cell factories and have been applied to the production of materials, fuels, commercial and industrial chemicals,

flavor, fragrances, pharmaceuticals, and foods.

Strain engineering efforts typically follow a cyclical approach consisting of 4 general steps: (1) Design (2) Build (3) Test (4) Learn. The design phase leverages current knowledge and computational approaches to rationally design a strain. Our systematic understanding of biology is incomplete due to its complexity; this renders rational strain design prohibitively difficult [3]. Introducing rationally designed changes into the DNA of a cell often results in a perturbed suboptimal metabolic or regulatory state, therefore providing an incomplete result [6]. Microbial cell factories additionally need to be robust to hard industrial conditions [6, 7]. These difficulties lead to multiple iterations of the strain engineering processes to converge on a feasible solution, which has been seen to require 6–8 years and over $50 million [3].

Bioengineers have a unique advantage over other engineering disciplines: biological systems have been solving challenges long before human intervention. The opportunity exists to leverage these built-in problem-solving processes to discover and elucidate biological function and generate application solutions. Adaptive Laboratory Evolution (ALE) methods are the formalization for leveraging the adaptive evolution process. When a biosynthetic pathway can be coupled with growth, ALE can provide adaptive mutations that will increase the throughput of this pathway. If a strain's health has been severely disrupted by a design change, ALE can serve to introduce mutations that rebalance the cell's homeostasis. ALE can additionally serve to harden strains against industrial conditions through tolerization [3, 6, 7].

Though ALE holds potential, it does have a barrier in the form of cost and complications. A full ALE experiment and sequencing costs just under $10,000 in consumables. ALE experiments run for approximately 1 month and each ALE replicate requires at least one passage a day, necessitating a trained employee. Additionally, complications often arise through contamination

and hypermutator strains.

A substantial amount of ALE data is being generated due to ALE's potential for discovery and application. The growth in ALE data has inspired multiple efforts in its consolidation and analysis [8, 9], though none with nucleotide-level resolution. In fields besides experimental evolution, the aggregation of pubic data is expected to enable discoveries not evident through single experiments [10]. Critical biomedical efforts, such as cancer and antimicrobial resistance research, have benefited from the meta-analysis of big data sets describing their respective fields.[11–13] Thus, aggregated public ALE data could be used to derive novel strain design principles and coinciding strains. The work of this dissertation seeks to aggregate public ALE data and implement methods to better find and understand causal mutations to design microbial strains with potential for application.

1.1 Consolidating experimental evolution mutational data

With new generations of high-throughput technologies becoming common in laboratories, enormous volumes of data are being made available through online databases. Mining of this type of data is an increasingly important part of biotech research; today's bioinformaticians are generating novel and impactful discoveries in biology through leveraging public datasets [10]. Databases of biological systems and components have been recognized as a critical resource for advancement and knowledge generation in biomedical fields [14]. There is a need for a consolidated and accessible experimental evolution mutation database. As already seen with many domains, a solution for this need is a web accessible platform that can report and export the aggregated ALE data. Meta-analysis of aggregated experimental evolution data could prove powerful in identifying and understanding novel adaptive mutations and insights into evolution.

1.2 Finding and understanding ALE key mutations

ALE experiments generate both neutral and beneficial mutations. Analysis of ALE experiments focuses on the beneficial mutations, therefore this subset of mutations must first be identified among all generated in the experiment. Mutations hypothesized to be adaptive, also known as *key mutations*, are first identified, then validated through the reintroduction into a wild-type host and performing fitness assays [15]. Once key and causal mutations have been found, work on understanding their benefit begins. Finding and understanding key mutations remains a primary challenge in the field of experimental evolution. The consolidation of ALE data and metadata along with rich mutation annotations have the potential to ease this burden. Consolidated ALE data could be used to statistically associate mutations to conditions, reducing the amount of conditions to consider for mutations sets and therefore deconvoluting the context of adaptive mutations. Rich mutation annotations could better inform what mutation effects, structural features, non-coding regulatory features, and cellular subsystems ALE mutations converge upon. These approaches enhance efforts in finding and understanding adaptive experimental evolution mutations.

1.3 Design, Build, and Test strains derived from ALE mutational data

Aggregated ALE mutational data and metadata represent an opportunity for new discoveries with experimental evolution. Mutational trends not evident within single experiments may become more clear with aggregated experiments. These trends may manifest for one or more specific experimental conditions on single genomics features or across a system of components

[16]. Meta-analysis of the aggregated ALE data using a combination of methods and biological knowledge-bases could reveal the principles underlying adaptive mutation trends. These principles could then be applied in strain designs towards industrially relevant objectives. New learning from these efforts will fill gaps in current biological knowledge and ultimately enhance strain design efforts.

1.4 References

1. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. en. *Nucleic Acids Res.* **44,** D67–72. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

2. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. en. *Science* **346,** 1258096. ISSN: 0036-8075, 1095-9203 (Nov. 2014).

3. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. en. *Cell* **164,** 1185–1197. ISSN: 0092-8674, 1097-4172 (Mar. 2016).

4. Garcia-Granados, R., Lerma-Escalera, J. A. & Morones-Ramirez, J. R. Metabolic Engineering and Synthetic Biology: Synergies, Future, and Challenges. en. *Front Bioeng Biotechnol* **7,** 36. ISSN: 2296-4185 (Mar. 2019).

5. Calero, P. & Nikel, P. I. Chasing bacterial chassis for metabolic engineering: a perspective review from classical to non-traditional microorganisms. en. *Microb. Biotechnol.* **12,** 98–124. ISSN: 1751-7915 (Jan. 2019).

6. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. en. *Nat. Biotechnol.* **33,** 1061–1072. ISSN: 1087-0156, 1546-1696 (Oct. 2015).

7. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. en. *Microb. Cell Fact.* **12,** 64. ISSN: 1475-2859 (July 2013).

8. Wang, X., Zorraquino, V., Kim, M., Tsoukalas, A. & Tagkopoulos, I. Predicting the evolution of Escherichia coli by a data-driven approach. en. *Nat. Commun.* **9,** 3562. ISSN: 2041-1723 (Sept. 2018).

9. Van den Bergh, B., Swings, T., Fauvart, M. & Michiels, J. Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. en. *Microbiol. Mol. Biol. Rev.* **82.** ISSN: 1092-2172, 1098-5557 (Sept. 2018).

10. Rudrapatna, V. A. & Butte, A. J. Open data informatics and data repurposing for IBD. en. *Nat. Rev. Gastroenterol. Hepatol.* **15,** 715–716. ISSN: 1759-5045, 1759-5053 (Dec. 2018).

11. Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortes-Ciriano, I., Zhou, D. C., Liang, W.-W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Van Allen, E. M., Liang, H., MC3 Working Group, Cancer Genome Atlas Research Network, Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R. & Ding, L. Comprehensive Characterization of Cancer Driver Genes and Mutations. en. *Cell* **173,** 371–385.e18. ISSN: 0092-8674, 1097-4172 (Apr. 2018).

12. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nat. Commun.* **9,** 4306. ISSN: 2041-1723 (Oct. 2018).

13. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. en. *Nat. Commun.* **5,** 4846. ISSN: 2041-1723 (Sept. 2014).

14. Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M. & Green, E. D. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. en. *J. Am. Med. Inform. Assoc.* **21,** 957–958. ISSN: 1067-5027, 1527-974X (Nov. 2014).

15. Palsson, B. Adaptive Laboratory Evolution This tool for studying molecular mechanisms and evolutionary dynamics follows bacteria as they adapt to environmental stress. *Microbe Wash. DC* **6,** 69. ISSN: 1558-7452 (2011).

16. Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D. & Gaut, B. S. The molecular diversity of adaptive convergence. en. *Science* **335,** 457–461. ISSN: 0036-8075, 1095-9203 (Jan. 2012).

Chapter 2

Consolidating Adaptive Laboratory Evolution Data into an Accessible and Computational Resource

2.1 Abstract

Adaptive Laboratory Evolution (ALE) has emerged as an experimental approach to discover causal mutations that confer desired phenotypic functions. ALE not only represents a controllable experimental approach to systematically discover genotype-phenotype relationships, but also allows for the revelation of the series of genetic alterations required to acquire the new phenotype. Numerous ALE studies have been published, providing a strong impetus for developing databases to warehouse experimental evolution information and make it retrievable for large-scale analysis. Here, the first step towards establishing this resource is presented: ALEdb

(http://aledb.org). This initial release contains over 11 000 mutations that have been discovered from eleven ALE publications. ALEdb (i) is a web-based platform that comprehensively reports on ALE acquired mutations and their conditions, (ii) reports key mutations using previously established trends, (iii) enables a search-driven workflow to enhance user mutation functional analysis through mutation cross-reference, (iv) allows exporting of mutation query results for custom analysis, (v) includes a bibliome describing the databased experiment publications and (vi) contains experimental evolution mutations from multiple model organisms. Thus, ALEdb is an informative platform which will become increasingly revealing as the number of reported ALE experiments and identified mutations continue to expand.

2.2 Background

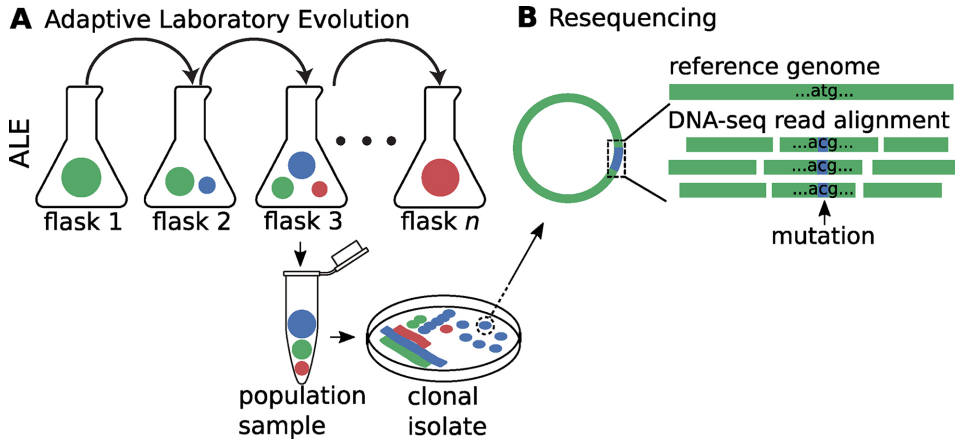


Figure 2.1: (A) An illustration of an ALE experiment where both a clonal and population sample are isolated from an intermediate (i.e., midpoint) flask. The Petri dish represents the streaking methodology for isolating a clonal colony from a population. (B) An illustration of how the resequencing process leverages a reference genome sequence and DNA-seq reads to identify mutations in an ALE sample.

Adaptive Laboratory Evolution (ALE) is a tool for the study of microbial adaptation. The

typical execution of an ALE experiment involves cultivating a population of microorganisms in defined conditions (i.e., in a laboratory) for a period of time that enables the selection of improved phenotypes. Standard model organisms, such as *Escherichia coli*, have proven well suited for ALE studies due to their ease of cultivation and storage, fast reproduction, well known genomes, and clear traceability of mutational events [1]. With the advent of accessible whole genome resequencing, associations can be made between selected phenotypes and genotypic mutations [2].

Beginning with a starting strain, an ALE experiment can be executed by serially passing a selected culture to a fresh flask of media (Figure 2.1A), enabling the strain passed to continue acquiring mutations under the experimental conditions without dilution of resources. Strains propagated during ALEs are assumed to be those that outcompeted their competition due to adaptive mutations. Additional methods to perform experimental evolutions have been reviewed [2, 3]. Whole genome comparative sequencing, or resequencing, is used to identify mutations within evolved strains relative to the evolution's starting strain (Figure 2.1B). ALE experiments can additionally involve replicate evolutions: identical evolutions that are often executed in parallel. Replicate ALEs can reveal the dynamics of adaptation by enabling research into converging genotypes within an experiment [4].

ALE methods have become important scientific tools in the study of evolutionary phenomena and have contributed to research in basic discovery and applied fields. Evolutionary biologists seek to examine the dynamics and repeatability of evolution and to better understand the relationship between genotypic and phenotypic changes [5]. ALE methods, along with the plummeting cost of sequencing, have greatly enabled their efforts, resulting in a variety of insights into adaptive evolution. ALE has often demonstrated that (i) increases in fitness diminish with

each new adaptive mutation [6], (ii) genotypic convergence through mutations can occur on the level of functional complexes [7] and (iii) interactions between mutations may cause nonlinear fitness effects [8].

ALE methods have also been leveraged in applications of synthetic biology to engineer microbes for commodity, industrial, and biopharmaceutical chemical synthesis [2]. Comprehensive whole genome rational design is rarely achievable due to the complexity of biological systems [4, 9]. The inability to provide for comprehensive solutions in genome engineering can result in strains which cannot maintain homeostasis, such as strains which cannot tolerate the concentrations of products they were designed to produce. ALE has been used to produce adaptive mutations that provide solutions for the gaps left by current rational genome engineering methods [10]. ALE can therefore complement rational genome engineering in the work to provide for a comprehensive whole genome solution to an application [2, 10].

Accurately interpreting the results of an ALE requires the identification of causal mutations within the given observed adaptations. Identifying causal mutations requires a clear understanding of the mechanistic effects of mutations on cellular components and systems. Due to the complexity of cellular systems, interpreting the effects of mutations has proven to be a primary challenge in ALE [4, 10]. A common approach to mutation functional analysis is a literature search on the mutation target (e.g. a given annotated ORF). Functional studies of genetic targets have traditionally served as primary resources for interpreting mutation effects, providing information on a sequence's biological function. Published ALE results can be used to enhance efforts to identify and understand new adaptive mutations. Researchers can work to understand their ALE mutations by considering published adaptive mutations in conditions similar to their own ALEs.

A review of ALE methods [2] lists 34 separate ALE studies. Each study reports on novel combinations of selection conditions and the resulting microbial adaptive strategies. Large scale analysis of ALE results from such consolidation efforts could be a powerful tool for identifying and understanding novel adaptive mutations.

A web platform named ALEdb (aledb.org) has been created to meet the need for accessible consolidated ALE mutations, conditions, and publication reporting. ALEdb additionally includes features to search for specific mutations, report key mutations, and export mutation data for custom analysis. With these features, ALEdb works to fill the gap in the field of experimental evolution for an accessible resource of consolidated experimental evolution mutations.

2.3 Results

2.3.1 A web platform to accelerate the work of ALE data to knowledge

The need for consolidated and accessible ALE experiment reporting has resulted in the generation of the web platform ALEdb (aledb.org). Eleven published ALE experiments, with a total of four distinct strains, 528 samples, and 11 792 observed mutations, serve as an initial data set (Figure 2.2).

Experimental evolution studies explore the solution space of a genome optimization problem through mutational events. This element of exploration has lead to a rich diversity of published ALE experimental conditions [2]. Those experimental conditions currently represented in ALEdb are genetic perturbations (11), stress inducing environments [11], different carbon sources [12–14], and evolution duration [5]. Strains can often adapt to these conditions with a variety of different evolutionary strategies, leading to different beneficial mutations. This leads to a

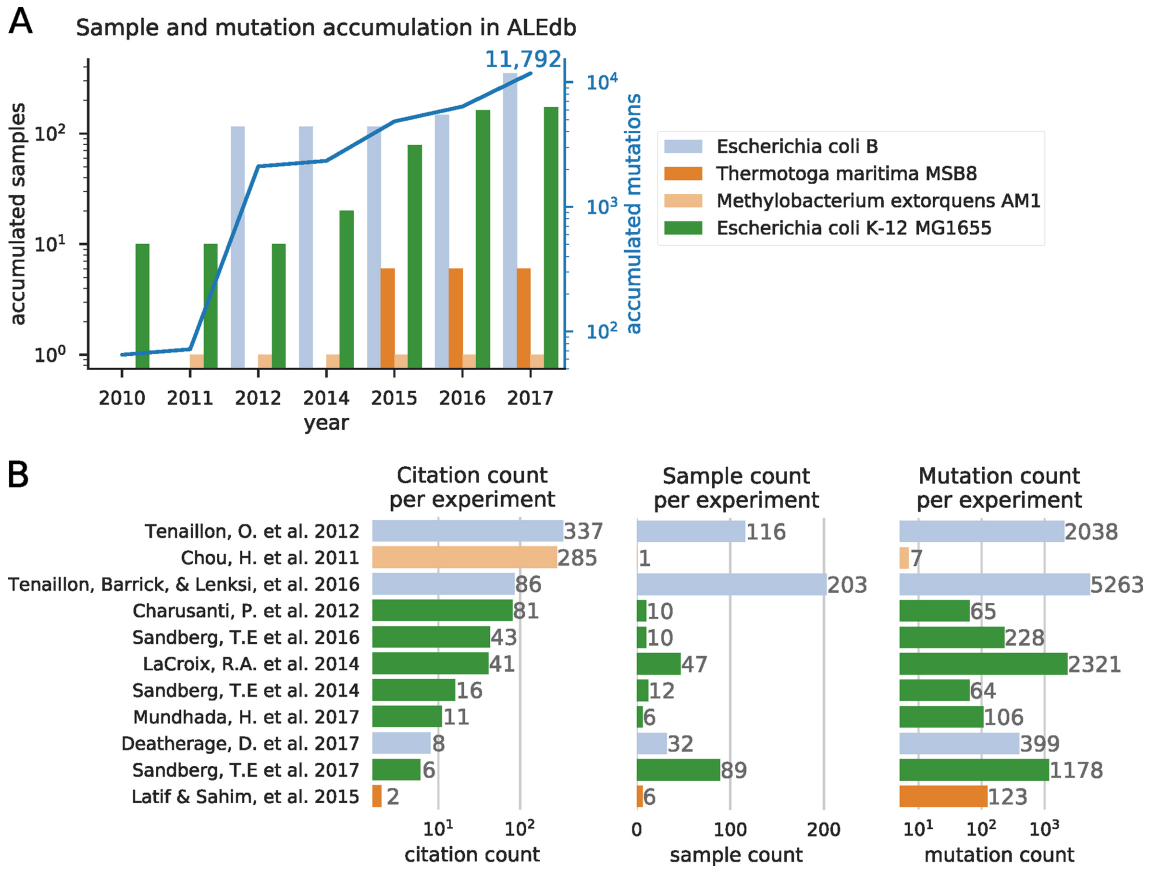


Figure 2.2: (A) A plot of the accumulated sequenced samples and mutations in ALEdb by publication year. (B) Each publication's sample and mutation contribution to ALEdb along with their citation count at the time of ALEdb's initial release. Citation counts were acquired from Google Scholar (scholar.google.com).

diversity in the mutations across ALE experiments. This rich variety of databased conditions and mutations have made ALEdb an attractive research resource, and further implementation has now made this information accessible through the web.

ALEdb's feature set was developed in response to the challenge of accessible ALE mutation reporting for an ALE experiment pipeline [15]. ALEdb's features enable intuitive navigation through consolidated ALE experiment data by providing two categories of features: those that describe individual ALE experiments, and those that describe all consolidated experiment data. To describe individual ALE experiments, ALEdb generates reports that detail the ALE muta-

tion lineages, key mutations, and experimental conditions per ALE sample. Each experiment is additionally described by a summary page which references the data's published work, includes summary visualizations, and annotates important details about the experiment dataset being hosted. To describe all consolidated ALE experiment data, ALEdb provides a mutation search feature, the ability to export the mutation data from one or more ALE experiments as spreadsheets, and an itemization of all publications that describe the databased mutations (Figure 2.3). ALEdb thus provides for a need in the experimental evolution community: a platform to search and explore consolidated experimental evolution mutation data.

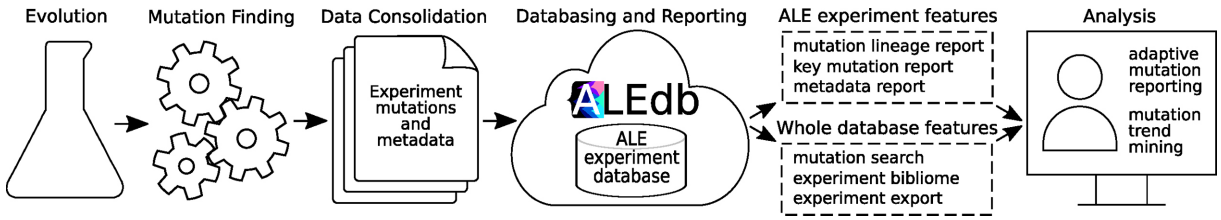


Figure 2.3: An illustration of the flow of ALE data from an experimental evolution to the generation of reports.

Mutation functional analysis is a major challenge in experimental evolution. Besides systems biology modeling methods, this task often involves searching the literature for similar results. The ALE mutations, conditions, and publications being consolidated into ALEdb can be leveraged in this work. ALEdb can enhance a user's mutation functional analysis by using a search-driven workflow to report if mutations similar to theirs have occurred in published ALE experiments. Through ALEdb's Search feature, users can query for mutations of interest using multiple descriptive parameters and become aware of any databased ALE experiments that manifest similar mutations. Knowing these experiments, users can review the conditions and key mutation reports which characterize their results and refer to their associated publications

through ALEdb's Bibliome page. These publications ultimately describe adaptive mutations and their functional analysis, which could be leveraged by users to better understand similar mutations in their own studies. ALEdb additionally includes the ability to Export mutation data for users interested in leveraging ALE data in applications beyond this platform (Figure 2.4). ALEdb's features are described in the following sections.

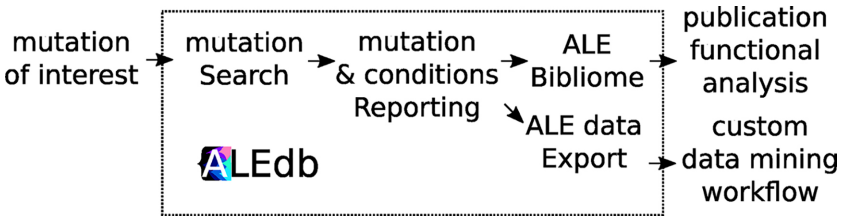


Figure 2.4: An illustration of the workflow for mutation functional analysis using ALEdb. Each step within the ALEdb group is the name of a user feature on the ALEdb platform.

2.3.2 Mutation search and reporting

ALEdb implements a mutation Search to enable users to quickly find mutations of interest. Search returns a report of mutations for all databased samples according to the following mutation descriptors: gene, genome position range, mutation type, sequence change, protein change, and experiment.

Mutation search, along with most other mutation reporting mechanisms on ALEdb, present their results in the form of mutation tables (Figure 2.5A). Each ALE experiment can be described as a series of mutation sets relative to an ALE's starting strain. Columns represent an ALE sample and are described with the experiment name, ALE (A#), flask (F#), isolate (I#) and technical replicate (R#) value to serialize samples (Figure 2.5B). Rows describe the specific mutation details that manifested within the samples, and values contained within cells

represent the allele frequency. Ordering sample columns from the earliest to latest sample in an ALE serves to render intuitive visualizations of temporal mutation trends. This format enables researchers to intuitively identify important mutational patterns, such as the fixed mutations within the corA gene and hns/tdk intergenic region (Figure 2.5A). Population samples will always be described by an isolate number (I#) of 0 and is the only sample type to carry allele frequencies less than 1.0. The information describing each mutation is generated by the mutation finding stage (Figure 2.2) and details a mutation's type, genetic target, and potential product effects. Mutation tables therefore describe the lineage of an ALE's final sample, or endpoint, according to the mutations that manifest during an evolution.

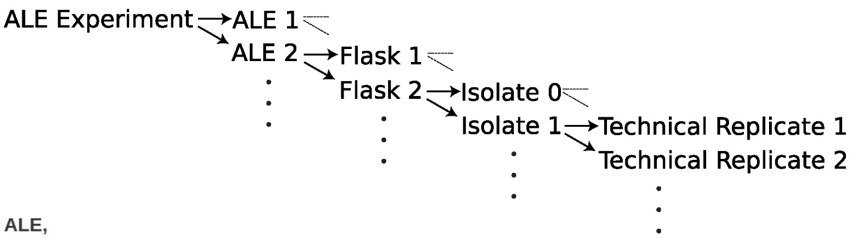
Researchers investigating ALE experiments require reporting that enables them to quickly understand which mutations are likely causal for adaptations; the mutation tables built by ALEdb are designed to meet this need. Among the many mutations that manifest within an ALE experiment, mutation rows that describe multiple alleles of a gene will cluster together according to their positions on the genome. This is illustrated with the corA mutation within Figure 2.5A. Due to the chronological sorting of the sample columns per ALE, a mutation that fixes across samples will manifest as an unbroken sequence of cells in a mutation row annotated with an allele frequency. This is illustrated with both the hns/tdk and corA mutations in Figure 2.5A. These two patterns are obvious to an observer and serve well to describe the adaptive mutational trends in ALE experiments.

ALE experiment mutations cannot be completely understood without considering the experiment's conditions. ALEdb includes reports that describe an experiment sample's strain, substrate, and environment (Figure 2.5C). These experiment Metadata reports can additionally be exported as spreadsheets for analysis external to ALEdb.

A

Position	Mutation Type	Sequence Change	Gene	Details	GLU A4 F66 I1 R1	GLU A4 F149 I1 R1	GLU A4 F237 I1 R1	GLU A4 F403 I0 R1	GLU A4 F403 I1 R1
1,551,659	SNP	G→A	adhP	P69S (CCA→TCA)		1.00			
3,998,893	DEL	Δ5 bp	corA	coding (220-224/951 nt)		1.00	1.00	1.00	1.00
3,999,402	DEL	(GGC)2→1	corA	coding (729-731/951 nt)	1.00				
1,292,256	MOB	IS1 (–) +8 bp	hns, tdk	intergenic (-110/-488)		1.00	1.00	1.00	1.00
3,203,742	SNP	G→A	ttdA	V11V (GTG→GTA)				0.21	

B



C

ALE, Flask, Isolate, Technical Replicate	Clonal or Population	Species	Strain	Additional Strain Details	Media	Substrate	Temperature
A4 F403 I0 R1	population	E. Coli	511145	BOP27	M9	Glucose(2g/L), NH4Cl(1g/L), KH2PO4(3g/L), O2	37.0
A4 F403 I1 R1	clonal	E. Coli	511145	BOP27	M9	Glucose(2g/L), NH4Cl(1g/L), KH2PO4(3g/L), O2	37.0

Figure 2.5: (A) An example mutation lineage report where samples are represented as columns, ordered from left to right as earliest to latest in an ALE. (B) An illustration of the information annotated by the sample column labels within mutation tables. In the case of the first sample column in this figure's mutation table, the label describes mutations from the first technical replicate, from the first isolate, from the sixty-sixth flask, from the fourth ALE, of the GLU experiment. (C) An ALE experiment metadata report.

2.3.3 Consolidated ALE knowledge

A key component in the utility of ALEdb is the per experiment knowledge built from the databased mutations. The Bibliome feature itemizes the publications that studied the ALE mutations databased within ALEdb. Users can leverage the mutation functional analysis within these publications toward understanding any similar mutations in their experimental evolutions.

2.3.4 ALE experiment mutation export

ALEdb implements an Export feature to give users the freedom to perform any analysis of interest on the hosted data. This feature enables users to extract one or more experiment mutation sets into comma separated value files. Users can then leverage custom analysis pipelines on the ALEdb's data towards generating novel results.

2.3.5 Automated ALE experiment key mutation reporting

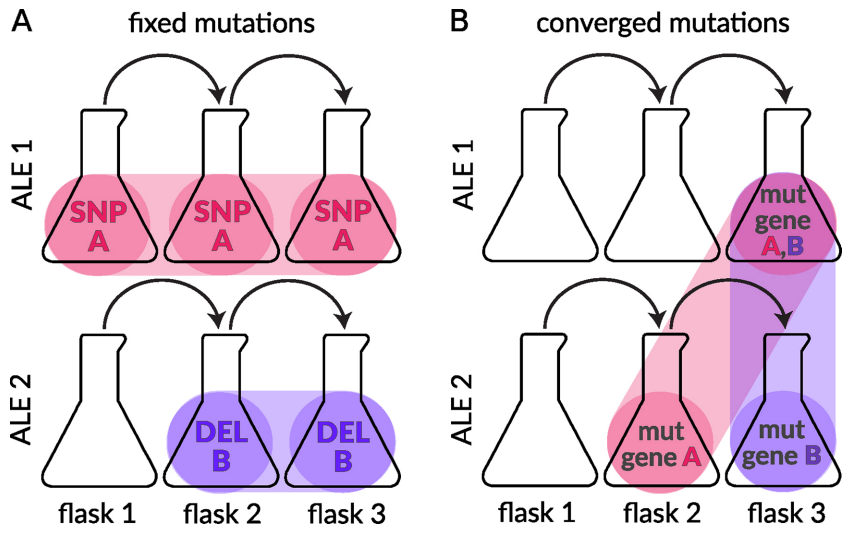


Figure 2.6: An illustration on the intuition utilized for fixed and converged mutation reports. (A) An illustration of fixed mutations. SNP A and DEL B occur in separate ALE replicates and persist through all subsequent flasks. (B) An illustration of converged mutations. Genetic targets A and B are seen to mutate across ALE replicates.

ALEdb includes features that automate the reporting of known ALE adaptive mutation trends. These trends are termed fixed and converged key mutations, where each trend describes a unique pattern of mutations occurring within or across multiple ALEs in an experiment. These patterns have been used in published ALE studies to identify adaptive mutations [11–13, 16]. The manual consolidation of adaptive mutation evidence can be prone to human error, inconsistent

between researchers, and time consuming. The automation of these common analyses contributes to more consistent analysis and more accurate results.

A fixed mutation is one in which first manifests in any ALE sample other than the endpoint, and is propagated to all following samples in the ALE. The propagation of a mutation from their emergence to an ALE's endpoint may describe the selection of a mutation due to its fitness benefits [12]. This analysis is only possible if an ALE experiment includes midpoint samples, providing the possibility of more than one data point per ALE mutation. The identification of fixed mutations is accomplished by organizing mutations according to the ALE's sample chronology and identifying mutations that emerge in a midpoint and manifest in all following samples of the same ALE (Figure 2.6A). ALEdb's fixed mutation reporting automates this analysis and reports results in the format described in (Figure 2.5A).

A converged mutation is one in which manifests in a genetic region seen to be mutated in multiple replicate ALEs (Figure 2.6B). This phenomenon describes evidence of a potential common adaptive trajectory between microbes exposed to the same conditions and has been leveraged in ALE analysis methods to more quickly identify mutations causal for adaptive phenotypes [12]. ALEdb's converged mutation reporting automates this analysis and reports results in the format described in (Figure 2.5A).

2.3.6 Design and implementation

ALEdb is implemented and deployed using a standard web application technology stack and a combination of user interface technologies. ALEdb's server-side hosts a MySQL database (https://www.mysql.com/), uses the Python based Django web framework (https://www.djangoproject.com/), and serves the content using Gunicorn

(http://gunicorn.org/) and Nginx (https://www.nginx.com/). ALEdb implements its user-interface with HTML, CSS, and JavaScript, along with a combination of Javascript libraries including Bootstrap (https://getbootstrap.com/), jQuery (https://jquery.com/), DataTables (https://datatables.net/), mutation-needle-plot (http://dx.doi.org/10.5281/zenodo.14561) and D3 (https://d3js.org/).

2.3.7 ALEdb mutation data usage case study

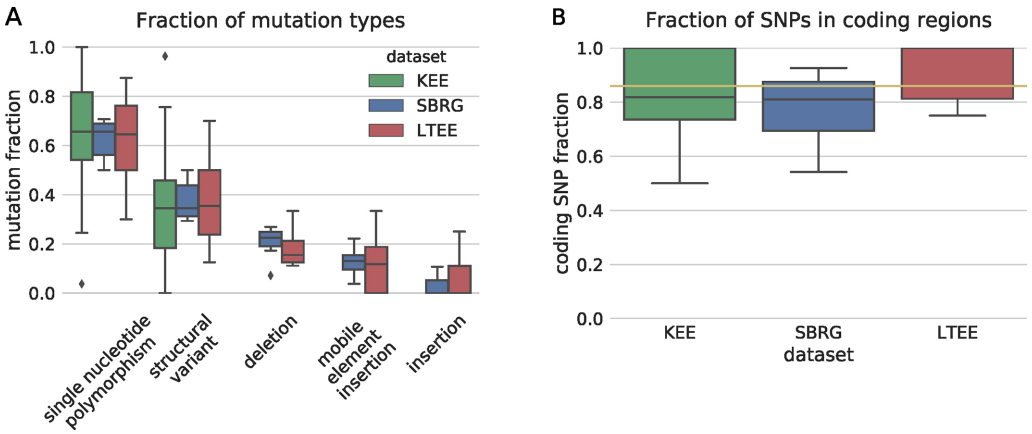


Figure 2.7: An example of multi-experiment analysis enabled by ALEdb's consolidated data. (A) Mutation type proportion distributions across endpoints for different sets of experiments. (B) Coding SNP proportion distributions across endpoints for different sets of experiments. The gold line represents the proportion of coding nucleotides within the E. coli bacterial genome (0.86) [17].

To demonstrate the potential for ALEdb as an experimental evolution mutation data resource, mutations from experiments databased within ALEdb were compared to mutation data published externally. *Escherichia coli* experimental evolution mutation data was gathered from three sources for comparison: (i) experiments executed by or with the Systems Biology Research Group (SBRG) [10–14, 16, 18], (ii) the Long Term Experimental Evolution (LTEE) [5] and (iii) the experimental evolution mutation set from the Dettman and Kassen consolidation studies

(KEE) [17]. Both the SBRG and LTEE datasets were found within ALEdb while the KEE dataset was annotated within its publication. The following results were generated external to ALEdb, where ALEdb data was extracted using the Export feature. The LTEE hypermutator strains were not included in this case study.

Mutation type distributions were calculated across experiment sets to compare their endpoint proportions. The mutation finding software used by the SBRG and LTEE [19, 20] describes mutations as single nucleotide polymorphisms (SNP), deletions (DEL), mobile element (MOB), and insertions (INS). The KEE set describes its mutations as SNPs or structural variants (SV), where SVs describe the combination of DEL, MOB and INS mutations. SNPs were the most common mutation across all datasets, with DELs, MOBs, and INS being the most common structural variants across SBRG and LTEE, in that order (Figure 2.7A). All experiment sets produced the same mutation type abundance and demonstrated similar distributions (Figure 2.7A, Table A.1).

The frequent manifestation of SNPs in experimental evolution endpoints has suggested that they play a role in adaptation. Previous work has investigated the selectivity of SNPs relative to the density of coding sequences within bacterial genomes [17]. It was proposed that if coding SNPs were more causal for adaptation than non-coding SNPs, their endpoint proportions would be significantly larger than the proportion of coding nucleotides within the E. coli bacterial genome (i.e. larger than 0.86) [17]. All of the coding SNP distributions for each of the three data sets overlap this average and do not provide evidence of being statistically different from each other (Figure 2.7B, Table A.2).

This case study serves as an example of the broad, multi-experiment analysis that can be readily performed with consolidated experimental evolution data. A goal of ALEdb is to provide

for this need through consolidating and reporting on experimental evolution mutation data.

2.4 Conclusion

ALEdb works to serve the current need for a mutation database in the field of experimental evolution. It is a platform designed for the integration and reporting of ALE mutation datasets and currently integrates the mutation data and published materials of eleven published ALE experiments. Additionally, multiple features are implemented within ALEdb to enable intuitive navigation and analysis. Finally, the case study included in this work demonstrates the potential for ALEdb as a mutation data resource for broad, multi-experiment analysis.

ALEdb will continue to be developed to meet the needs of consolidating, reporting, and navigating ALE experiment data. This initial release of ALEdb considers previously generated mutation datasets. ALEdb will continue to grow with future inclusion of published ALE experiment results from currently contributing and new research organizations.

Acknowledgements

The authors would like to thank Lars Schoning, Kai Chen, and Pankaj Kumar Singh for their work on the infrastructure necessary to release ALEdb to the public. The authors would like to thank Dr. Ali Ebrahim, Dr. Ryan LaCroix, Kaiwen Zhang and Dylan Skola; their initial designs provided the foundation for ALEdb.

This work and the open access charge was supported by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark [NNF10CC1016517] and by the National Institutes of Health's National Institute of Allergy

and Infectious Diseases [U01AI124316].

Chapter 2 is a reprint of material published in: **Patrick V Phaneuf**, Patrick V Phaneuf, Dennis Gosting, Bernhard O Palsson, Adam M Feist, "ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation." Nucleic Acids Res. 2019 Jan 8;47(D1):D1164–71. The dissertation author is the primary author.

2.5 References

1. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. en. *Nat. Rev. Genet.* **4,** 457–469 (June 2003).

2. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. en. *Microb. Cell Fact.* **12,** 64. ISSN: 1475-2859 (July 2013).

3. Rabbers, I., van Heerden, J. H., Nordholt, N., Bachmann, H., Teusink, B. & Bruggeman, F. J. Metabolism at evolutionary optimal States. en. *Metabolites* **5,** 311–343 (June 2015).

4. Palsson, B. Adaptive Laboratory Evolution This tool for studying molecular mechanisms and evolutionary dynamics follows bacteria as they adapt to environmental stress. *Microbe Wash. DC* **6,** 69. ISSN: 1558-7452 (2011).

5. Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Medigue, C., Schneider, D. & Lenski, R. E. Tempo and mode of genome evolution in a 50,000-generation experiment. en. *Nature* **536,** 165–170 (Aug. 2016).

6. Barrick, J. E. & Lenski, R. E. Genome-wide mutational diversity in an evolving population of Escherichia coli. en. *Cold Spring Harb. Symp. Quant. Biol.* **74,** 119–129 (Sept. 2009).

7. Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D. & Gaut, B. S. The molecular diversity of adaptive convergence. en. *Science* **335,** 457–461. ISSN: 0036-8075, 1095-9203 (Jan. 2012).

8. Wunsche, A., Dinh, D. M., Satterwhite, R. S., Arenas, C. D., Stoebel, D. M. & Cooper, T. F. Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. en. *Nat Ecol Evol* **1,** 61 (Mar. 2017).

9. Williams, T. C., Pretorius, I. S. & Paulsen, I. T. Synthetic Evolution of Metabolic Productivity Using Biosensors. en. *Trends Biotechnol.* **34,** 371–381 (May 2016).

10. Mundhada, H., Seoane, J. M., Schneider, K., Koza, A., Christensen, H. B., Klein, T., Phaneuf, P. V., Herrgard, M., Feist, A. M. & Nielsen, A. T. Increased production of L-serine

in Escherichia coli through Adaptive Laboratory Evolution. en. *Metab. Eng.* **39,** 141–150 (Jan. 2017).

11. Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M. & Feist, A. M. Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. en. *Mol. Biol. Evol.* **31,** 2647–2662 (Oct. 2014).

12. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Appl. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

13. Sandberg, T. E., Long, C. P., Gonzalez, J. E., Feist, A. M., *et al.* Evolution of E. coli on [U-13C] glucose reveals a negligible isotopic influence on metabolism and physiology. *PLoS* (2016).

14. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Appl. Environ. Microbiol.* **83** (July 2017).

15. LaCroix, R. A., Palsson, B. O. & Feist, A. M. A Model for Designing Adaptive Laboratory Evolution Experiments. en. *Appl. Environ. Microbiol.* **83** (Apr. 2017).

16. Charusanti, P., Conrad, T. M., Knight, E. M., Venkataraman, K., Fong, N. L., Xie, B., Gao, Y. & Palsson, B. O. Genetic Basis of Growth Adaptation of Escherichia coli after Deletion of pgi, a Major Metabolic Gene. *PLoS Genet.* **6,** e1001186 (Nov. 2010).

17. Dettman, J. R., Rodrigue, N., Melnyk, A. H., Wong, A., Bailey, S. F. & Kassen, R. Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. en. *Mol. Ecol.* **21,** 2058–2077 (May 2012).

18. Long, C. P., Gonzalez, J. E., Feist, A. M., Palsson, B. O. & Antoniewicz, M. R. Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 222–227 (Jan. 2018).

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

20. Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., Knoester, D. B., Reba, A. & Meyer, A. G. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. en. *BMC Genomics* **15,** 1039 (Nov. 2014).

Chapter 3

Causal mutations from adaptive laboratory evolution are outlined by convergence on multiple scales of genome annotations and condition-specificity

3.1 Abstract

3.1.1 Background

Adaptive Laboratory Evolution (ALE) has emerged as an experimental approach to discover mutations that confer phenotypic functions of interest. However, the task of finding and

understanding all beneficial mutations of an ALE experiment remains an open challenge for the field. To provide for better results than traditional methods of ALE mutation analysis, this work applied enrichment methods to mutations described by a multiscale annotation framework and a consolidated set of ALE experiment conditions. A total of 25,321 unique genome annotations from various sources were leveraged to describe multiple scales of mutated features in a set of 35 *Escherichia coli* based ALE experiments. These experiments totalled 208 independent evolutions and 2641 mutations. Additionally, mutated features were statistically associated across a total of 43 unique experimental conditions to aid in deconvoluting mutation selection pressures.

3.1.2 Results

Identifying potentially beneficial, or key, mutations was enhanced by seeking coding and non-coding genome features significantly enriched by mutations across multiple ALE replicates and scales of genome annotations. The median proportion of ALE experiment key mutations increased from 62%, with only small coding and non-coding features, to 71% with larger aggregate features. Understanding key mutations was enhanced by considering the functions of broader annotation types and the significantly associated conditions for key mutated features. The approaches developed here were used to find and characterize novel key mutations in two ALE experiments: one previously unpublished with *Escherichia coli* grown on glycerol as a carbon source and one previously published with *Escherichia coli* tolerized to high concentrations of L-serine.

3.1.3 Conclusions

The emergent adaptive strategies represented by sets of ALE mutations became more clear upon observing the aggregation of mutated features across small to large scale genome annotations. The clarification of mutation selection pressures among the many experimental conditions also helped bring these strategies to light. This work demonstrates how multiscale genome annotation frameworks and data-driven methods can help better characterize ALE mutations, and thus help elucidate the genotype-to-phenotype relationship of the studied organism.

3.2 Background

Adaptive Laboratory Evolution (ALE) is used to study microbial populations under specific conditions over many generations and provides insights into the underlying mechanisms of adaptive phenotypes. Mutations observed from ALE experiments have proven valuable for both biological discovery and applied biotechnology, such as the elucidation of the rate and mechanisms of mutation development [1–5] and the design of industrially relevant strains for increased bioproduction [6]. The increased scale of ALE experiments—due to the low cost of sequencing and the inclusion of intermediate/midpoint samples [7], multiple replicates [8], and population samples [9]—has increased the number of mutations that require analysis. While the identification of "key mutations", or those mutations hypothesized as being adaptive, is better enabled with more mutation data [8], traditional methods are not well suited for large scale sets of ALE mutations. The potential of mutated non-coding regulatory features contributing to an adaptive phenotype further complicates the set of features to consider when seeking to understand ALE mutations. Moreover, multiple experimental parameters can contribute to the selection pressure that an organism experiences in experimental evolution [6, 10]. Ultimately, the primary

challenges with traditional mutation functional analysis are finding the subset of adaptive mutations among the many that emerge during an ALE experiment and understanding the adaptive mechanisms of these mutations relative to specific selection pressures.

The main concern with traditional ALE mutation analysis is the mutated genes and how the sequence changes affect their function. Identifying commonly mutated genetic features (genes or intergenic regions) across replicate ALEs, known as convergence, has been established as a primary method for identifying potentially causal mutations, or key mutations, in ALE experiments [8]. Mutation convergence on broader levels of genomic organization has provided evidence that mutations targeting different features can accomplish similar adaptive functional changes [11]. This bottom-up convergence of mutated features across multiple scales of annotations enables a top-down approach to understand large sets of mutations: researchers can consider the broader functional annotations emphasized by large sets of small mutated features before analyzing individual mutations. Enrichment methods have been developed to identify over-represented classes among large collections [12]. Thus, enrichment methods can leverage high-throughput genome-wide data and molecular biology ontologies to identify enriched biological functions from large sets of mutated genes. However, the challenges of examining non-coding regulatory features and deconvoluting selection pressures for ALE mutations remain.

The accumulating wealth of information on the molecular biology of *Escherichia coli* K-12 MG1655 has led to the emergence of knowledge and data resources that can help solve challenges in understanding ALE results. Genome annotation frameworks such as regulons [13], pathways [14], and clusters of orthologous groups (COG) [15] describe functionally related coding and non-coding regulatory features on multiple scales of genome annotation. Similar to gene set enrichment analysis [12], significant enrichment can be investigated across multiple scales of

genome annotations for meaningful convergence events. Additionally, the increased amount and scale of ALE experiments have led to efforts in consolidating their results. ALEdb, a web-based platform, reports on the mutations and experimental conditions from multiple experimental evolutions [16]. The mutations and conditions found in ALEdb can be used to associate mutated features to conditions and provide evidence on the selection pressures for ALE mutations.

Here, we address the challenges with finding and understanding adaptive mutations through two approaches. The first is to better identify key mutations than traditional means by seeking statistically significant mutation convergence across multiple scales of genome annotations. The second is to better understand these key mutations through their enrichment of functional annotations and their statistical associations to experimental conditions. We anticipate that the approaches described here will provide the ability to deconvolute systematic targets of adaptive mutations and their selection pressures to aid in improving ALE mutation functional analysis.

3.3 Results

3.3.1 A framework for finding key mutations using significant convergence on multiple scales of genome annotations

To comprehensively characterize the frequency of mutated features on the genome for a set of ALE experiments, mutation annotations should consider the variety of non-coding regulatory features along with coding features. Genome references typically include gene annotations, enabling mutation calling pipelines to describe mutations as affecting genes and/or intergenic regions, though there exists a multitude of additional feature and functional annotations. In this

work, multiple annotation types were used to consider different scales of mutated features (i.e., gene, operon, regulon, etc.) for the *E. coli* K-12 MG1655 genome (Figure 3.1a), with a total of 25,321 unique genome annotations for mutations (Figure 3.2a). The smallest-scale genome annotation type used in this work was referred to as genomic features and described coding features, non-coding regulatory features, and non-coding intergenic regions of unknown function. The regulatory features considered were transcription factor binding sites (TFBS), promoters, terminators, attenuator-terminators, and ribosome binding sites (RBS). These small regulatory features and genes were described by RegulonDB version 10.0 [17]. Mutated coding and non-coding features were then mapped to their transcription units (TU), then operons and functional annotations (Figure 3.1b). The sources of functional annotations used in this work were pathways as described by the PATRIC database [14], regulons of RegulonDB version 10.0 [17], and clusters of orthologous groups (COGs) [15]. This multiscale annotation framework included a level of annotation with only genes and intergenic regions to provide the expected evidence of convergence according to previously established methods of finding key mutations [8].

The Sankey diagrams [18] visualizations used in this work demonstrate the number of mutations to each feature and the connectivity of smaller-scale features to their larger-scale counterparts (Figure 3.1c). The underlying data structure is a directed acyclic graph (DAG). The DAG represents the frequency of mutated genomic features and how they aggregate towards broader annotation type features. The nodes in the DAG represent a mutated feature. The edges of the DAG represent the connection between those features, where their direction typically goes from smaller to broader annotation types (Figure 3.1b). The edge weights represent the accumulating instances of mutated genomic features across annotation scales. The node weights represent the sum of incoming edge weights. The visualization presented in this study for the

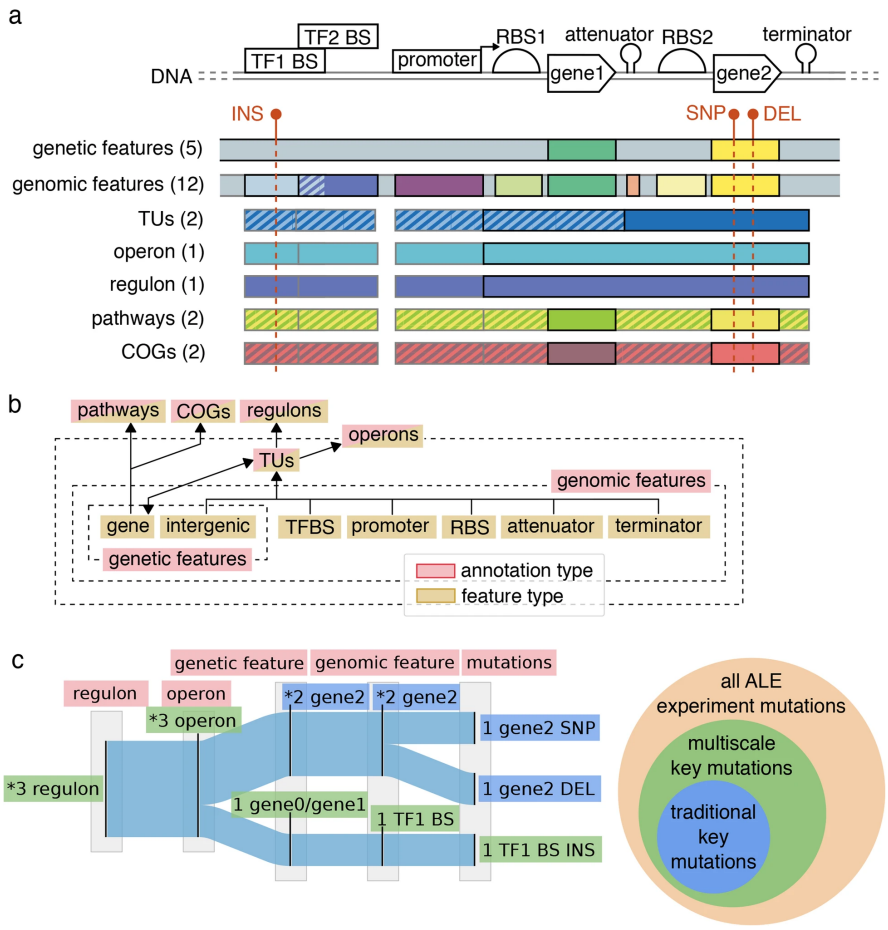


Figure 3.1: An illustration of how features were connected together in this work's multiscale annotation framework. (a) An illustration of the variety of annotations, given in parenthesis per annotation type, for the *E. coli* K-12 MG1655 genome and how a single mutation can affect multiple features across different scales of annotations. The diagonal striped regions illustrate overlapping features. Dark-edged rectangles represent defining features for an annotation type. Grey-edged boxes represent operational regions or features associated with the defining features. Features are not to scale. (b) A flow diagram demonstrating the mapping of mutated small-scale features onto larger-scale features of the multiscale annotation framework of *E. coli* K-12 MG1655 from this work. (c) An example of the Sankey diagram visualization used in this work to demonstrate the number of mutations to each feature and the connectivity of smaller-scale features to their larger-scale counterparts. Each feature is annotated with a value representing the number of instances the feature was observed to be mutated. Significantly enriched features are annotated with an asterisk (*). Mutated features contributing to the significant enrichment of higher-level annotations are considered as hosting key mutations. The Venn diagram illustrates the potential for finding more key mutations than traditional methods through multiscale scale annotation mutation convergence.

DAG additionally includes the different mutation types affecting each genomic feature. A DAG is constructed per mutation set, where the mutated genomic features are first established from the mutation information. A single mutation may introduce multiple genomic features as well as multiple mutations may only contribute to the mutation frequency of one genomic feature. If no explicit genomic feature can be connected to a mutation, an intergenic region annotation is assigned according to flanking genes. Transcription units (TU) are part of the DAG, though aren't included in this study's Sankey diagrams due to containing mostly redundant mutated feature convergence with operons. TUs connected genes to higher-level annotations associated with regulatory mechanisms. TUs also enabled the connection between non-coding features and gene-based functional annotations, such as pathways and COGs, according to the genes hosted on a TU (Figure 3.1b). Connections between mutated genomic features and broader annotations mostly rely on the relationships established within the multiple sources of operational and functional annotations integrated for this study. Mutated intergenic regions of unknown function are assigned TUs according to overlapping nucleotide positions. After TUs are assigned, functional annotations can be connected in the same manner as with mutated genomic features of known functions.

Large features are aggregations of many smaller features and consequently manifest mutation convergence more easily by random chance (Figure 3.2b). A statistical enrichment method was applied to quantify the significance of mutation convergence on features of the genome and prioritize their importance for functional analysis. The method assumed that each nucleotide in the genome has the same probability of being spontaneously mutated in an ALE. This assumption translated to annotated features on the genome having a probability of being mutated proportional to their length. Though this assumption may not perfectly reflect the distribution of

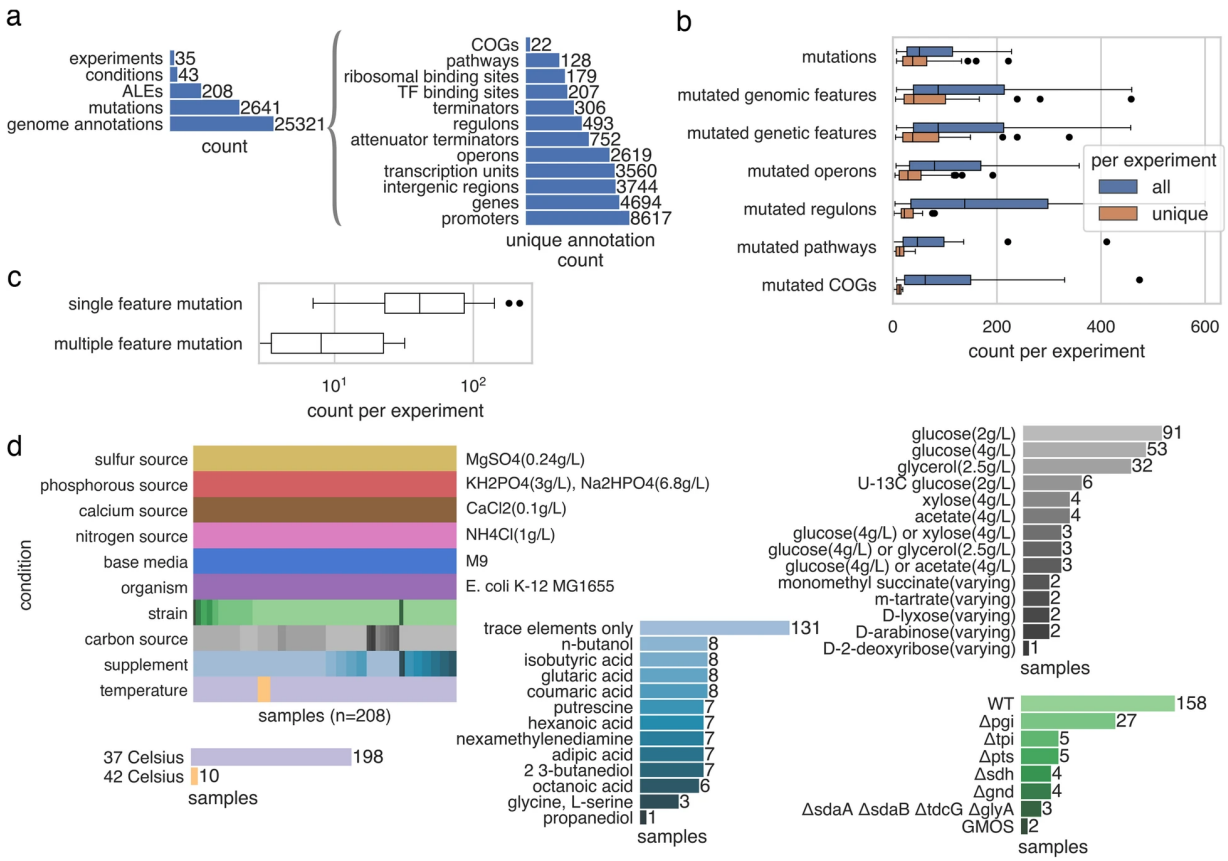


Figure 3.2: Magnitudes for different aspects of the ALE experiment mutations, conditions, and annotations used within this study. (a) The total number of experiments, ALEs, mutations, conditions, and unique genome annotations. (b) The distribution of mutations and annotation type features mutated per experiment. (c) The distribution of single versus multiple feature mutations per experiment, based on genomic features. (d) The conditions describing ALE experiments from ALEdb, the combinations of condition labels across ALEs, and the amount of each specific label for conditions with multiple labels.

mutations across a genome, it has been experimentally validated to represent their general distribution with mutation accumulation studies [18]. Thus, studies searching for signals of mutation selection commonly use a random distribution of mutations across the amino acids or nucleotides for a set of features [19] or a whole genome [20] as a null hypothesis in their statistical enrichment tests. Significant enrichment of annotated features was tested separately per annotation type (genomic features, genetic features, operons, etc). For annotation types that don't have

explicit coverage of the entire genome, an additional feature was added to the annotation type set that represented these non-annotated regions. For each annotation type, a permutation test of 10,000 iterations was executed where mutations were distributed across a specific annotation type's features using their mutation probabilities per iteration [19]. The length of each regulon was defined as the total number of nucleotides in the TUs considered within each regulon. The length of each COG or pathway was defined as the total number of nucleotides in the TUs in which the gene connected to the COG or pathway was found, excluding the lengths of genes within the TU not connected to the specific COG or pathway. Features with more than one mutation and a permutation test p-value<0.05 (Bonferroni corrected) were considered to be significantly enriched by mutations. A mutated feature found to be significantly enriched had an asterisk (*) prepended to its label within the flow diagram (Figure 3.1c). Finally, all mutations contributing through convergence to the significance of an enriched feature were considered key mutations (Figure 3.1c).

3.3.2 The amount and diversity of ALE mutated feature types and experiment conditions

The dataset used within this work contained 35 *Escherichia coli* K-12 MG1655 based ALE experiments from ALEdb [16], totaling 208 independent evolutions and 2641 mutations (Figure 3.2a). Within this dataset, experiments have a median of 51 mutations, with a median of 38 being unique (Figure 3.2b). As broader annotations types are considered, a smaller amount of unique features are mutated per ALE experiment (Figure 3.2b). Multi-nucleotide structural variants or overlapping features (Figure 3.1a) on the genome can result in more than one genomic feature affected by a mutation (Figure 3.2c), especially in the case of the numerous small

regulatory features (Figure 3.2a), therefore leading to more mutated features than mutations in an ALE experiment. Some of these small non-coding features can additionally regulate more than one operon, further complicating the analysis of their effects on the host. These types of mutations and relationships between features increase the number of artifacts to consider with the functional analysis of ALE mutations. In this dataset, mutations affected a median of 87 genomic features per experiment (Figure 3.2b), demonstrating that there could be more mutated features in an ALE experiment than individual mutation events.

This dataset tracked 10 different types of experimental condition types with a total of 43 unique conditions (Figure 3.2a, d), describing the organism and environment of the ALE experimental design. Annotations affected by mutations were statistically associated with these conditions to better understand which may have selected for the ALE mutations. Fisher's Exact Test was used for the statistical association, where a condition and mutated feature were considered associated if their odds ratio was greater than 1 with p-value$<$0.01 (Bonferroni corrected). A p-value$<$0.01 was used to measure significance rather than a p-value$<$0.05 to reduce the amount of false-positives with associations.

3.3.3 Applying methods across a large consolidated ALE mutation dataset and to individual case studies

The impact of annotating mutations using multiscale genomic features was explored using 18 *E. coli* K-12 MG1655 based ALE experiments. We first present a small case study from a Δpgi ALE experiment [**Charusanti**, 21] to demonstrate how the mutation of both coding and non-coding features can accomplish the same fitness benefit. Further, the abundance of both coding and non-coding features, along with their contributions to key mutation identification

across the data set, was investigated. We then examined the impact of these methods on the number of key mutations found across these ALE experiments. Finally, we applied these methods to the mutation sets of two ALE experiments separately, to demonstrate the novel value of the new types of evidence with mutation functional analysis. The first ALE experiment represents a newly released mutation set from a previous ALE growth-rate characterization study to growth on glycerol as a carbon source [22]. The second ALE experiment represents a previously published mutation set from an L-serine tolerance ALE [23] with a newly outlined adaptive strategy based on evidence from this work's methods.

Key mutations increase through the bottom-up convergence of mutated features on a multiscaled annotation framework

The convergence of mutated coding and non-coding regulatory features onto broader annotations was investigated with existing ALE experiment mutations. Non-coding mutations have previously been seen to play a beneficial role in ALE phenotypes [21, 24]. To illustrate the convergence of both coding and non-coding mutations onto a broader scale of annotation, a case study around operons involved in transhydrogenase activity is presented using data from one of the ALE experiments consolidated in this study's data set. This particular ALE experiment reported on the adaptation to a *pgi* knockout (i.e., Δpgi), finding that the proteins PntA and PntB were rendered non-functional through truncations [21]. The study additionally observed a mobile insertion element mutation upstream of the *pntA* gene in a replicate lacking *pntA* and *pntB* mutations. The same ALE study had seen SNPs to both *sthA* and its upstream regions. The mutations to these upstream regions were hypothesized to change the expression of the downstream genes in a way that benefited the Δpgi host. Advanced phenotypic characterization

of the endpoint strains with *pntA*, *pntB*, *sthA*, and upstream mutations revealed a convergence to equivalent intracellular states, demonstrating the same phenotype achieved by different mutations. The potential impact of the upstream mutations became more evident when integrating annotations for non-coding regulatory features and observing the convergence of all small mutated features onto broader annotations such as operons (Figure 3.3a). The upstream mutations targeted promoters for operons of the mutated *pntA*, *pntB*, and *sthA* genes. The convergence of mutations onto the *pntAB* and *sthA* operons highlighted their importance in this adaptation and serve to contextualize each mutation involved.

The abundance of regulatory features with key mutations was investigated with a set of *E. coli* K-12 MG1655 ALE experiments. A variety of non-coding regulatory features alongside genes was integrated to reduce the proportion of non-coding mutation targets with unknown function (Figure 3.3b). It was observed that all small non-coding regulatory features had mutations involved in statistically significant convergence (p-value<0.05, permutation test, mutations to feature>1). Except for promoters, the median proportions of non-coding features involved in significant convergence across ALE experiments were higher than that of coding features (Figure 3.3c). In fact, in all cases, mutated RBS were always seen to be involved in statistically significant convergence.

The convergence of small mutated features onto broader annotations and functions may not always be straightforward to manually identify with traditional methods. For example, an L-serine tolerance ALE experiment acquired adaptive mutations to *rho* or the *trxA/rhoL* intergenic region in different independent replicates [23]. The intergenic mutation targeted the *rhoLp* promoter for the operon hosting both *rhoL* and *rho* genes, demonstrating that beneficial mutations may be found in features not immediately adjacent to a given gene. Further, beneficially

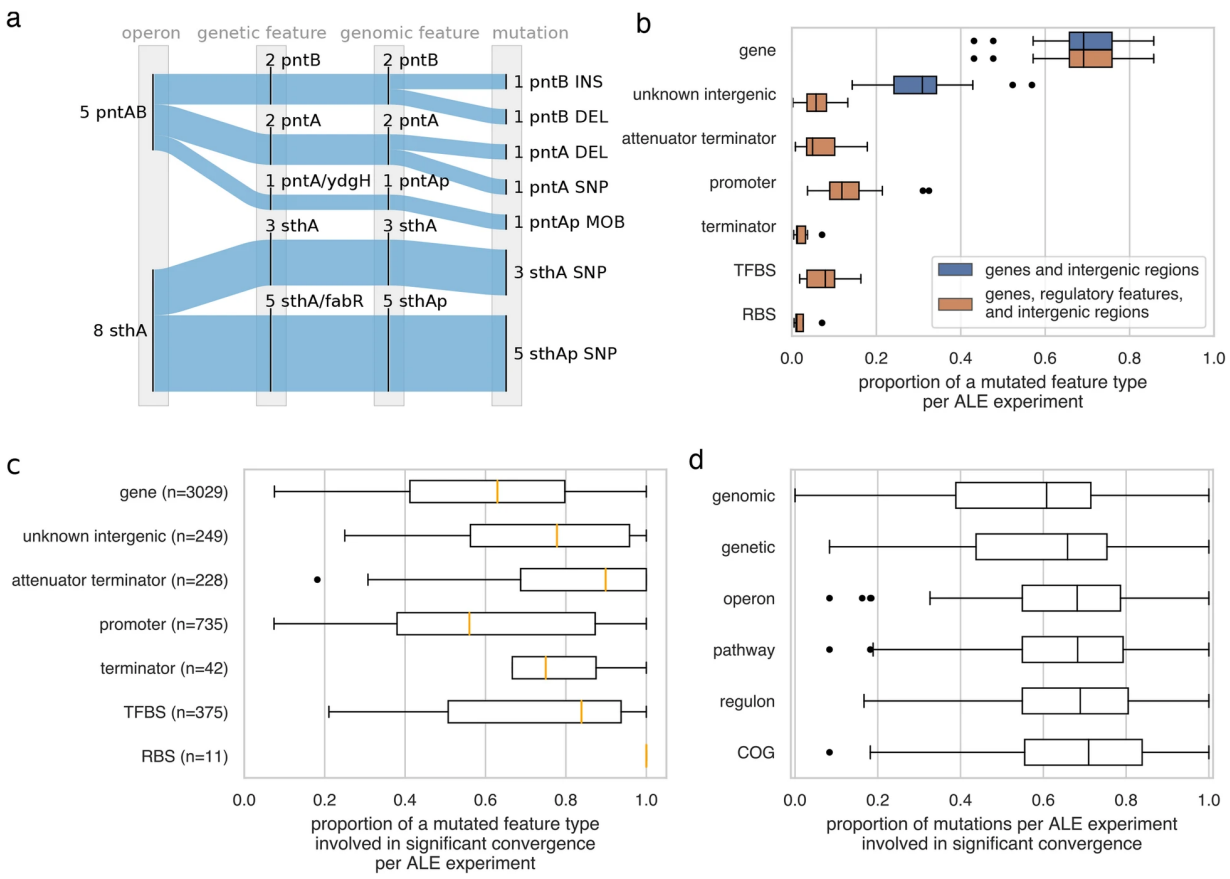


Figure 3.3: The amount of identified key mutations increase through the bottom-up convergence of mutated features on a multiscaled annotation framework. (a) The convergence of mutated genes and regulatory features onto *pntAB* and *sthA* operons from a Δ*pgi* ALE experiment on *E. coli* K-12 MG1655 [21, 25]. (b) Proportions of mutated features according to different genomic feature types. The distributions were generated by finding the proportion of each mutated feature per ALE experiment. (c) The proportions of mutated feature types across ALE experiments that were involved in significant convergence (textitp-value<0.05, permutation test, mutations to feature>1) on the multiscale annotation framework. The number of times each feature type was observed mutated is included ('n'). (d) Proportion of mutated features across ALE experiments that had significant convergence up to an annotation type on the multiscale annotation framework of Figure 3.1.

mutated features belonging to the same system are not always found on the same operon. For example, in an ALE experiment that resulted in a strain which had a elevated levels of aromatic amino acids, mutations to the *rcsA*, *rcsB*, and *yrfF* genes were found, along with mutations to *fliR/rcsA* and *nudE/yrfF* intergenic regions [26]. Each of these key mutated genes are hosted

on different operons and their mutations either truncated their coding sequences or repressed their expression via mutations to their promoters and other non-coding features. All three genes belonged to the Rcs stress response system, whose activation by a perturbation (i.e., a *ptsHIcrr* knock out) in the starting strain was detrimental to population growth. The deactivation of the Rcs system through the mutations distributed across multiple operons was a key feature in the genotype which ultimately enabled the heightened aromatic amino acid levels. Though these types of mutation convergences can be manually identified with the prerequisite knowledge and detailed annotation, the growth in mutation datasets and the use of more complex organisms render their identification less likely.

It was observed that the addition of broader types of annotations to the multiscale annotation framework generally increases the number of key mutations found within ALE experiments (Figure 3.3d). This is demonstrated by the increase in the median proportion of significantly convergent mutations (p-value<0.05, permutation test, mutations to feature>1) in ALE experiments with each broader annotation type. For example, 62% (median) of the mutated features annotated only with genomic features (i.e., coding and non-coding features) were found to be significant. When considering the COGs for these small features, 71% (median) of mutated features were involved in significant convergence.

Top-down functional analysis and mutated feature-to-condition associations for *E. coli* K-12 MG1655 growth selection on glycerol highlights changes to carbon catabolism and its repression

The methods of top-down functional analysis and mutated feature-to-condition association analysis contextualized mutations in an ALE experiment evolving *E. coli* K-12 MG1655

with some variants [27] on glycerol as a carbon source and resulted in an understanding of key mutations. The ALE experiment was previously executed with 30 independent replicates to test the effects of passage volume on endpoint adaptive phenotypes and represents a relatively large scale ALE experiment [8]. Besides passage volumes, the conditions used were similar to the earlier glycerol evolution of Herring *et al.* 2006 [28]. The replicate ALE endpoints were represented by a total of 51 samples (24 clonal, 27 population). Of the 51 total endpoints, 18 were represented with both clonal and population samples, while 6 endpoints were represented with one or more population samples, and 6 were represented with one or more clonal samples. Comprehensive whole-genome resequencing was performed for this work. Only mutations with a sample frequency of 50% or more were considered for population samples and those mutations that overlapped between endpoint samples were only considered once (see Methods). Collectively, 148 mutations were analyzed, representing a large ALE experiment mutation set for manual mutation functional analysis.

The CRP regulon hosted 62 mutated features, the largest amount within this ALE experiment. The CRP regulon describes the functions associated with the dual regulator CRP and the cAMP receptor protein, where cAMP is known as the catabolite gene activator protein. CRP is known to regulate many functions, one of which is carbon catabolite repression (CCR) [29, 30] and represses the metabolism of carbon sources besides glucose. The *glpK*, *cyaA*, and *crr* genes were the three most often mutated in this experiment, hosting 28, 21, and 10 mutations, respectively, and are within the CRP regulon (Figure 3.4a). The CRP regulon and genes were significantly converged upon by mutations (all p-value<0.001, permutation test, mutations to feature>1) and mutations to these genes were strongly associated with the selection pressure of glycerol as a carbon source (p-value<0.001, Fisher's exact test, Figure 3.4b). In a similar study

[28], the *glpK*, *crr*, and *cyaA* genes were also identified as being key to adaptation on glycerol as a carbon source. These associations, along with their convergence on the CRP regulon, suggest the mutations to *glpK*, *crr*, and *cyaA* as likely selected by the pressure to increase growth rates by rapidly utilizing glycerol as a carbon source.

The significant convergence of mutated features on functional annotations contextualized the targets of the *glpFKX* operon mutations. The significantly convergent *glpFKX* operon (p-value<0.001, permutation test, mutations to feature>1) hosted 30 instances of mutated features, corresponding to 28 coding SNPs in *glpK*, where two of these mutate a GlpR TFBS (located inside the coding sequence of *glpK*). These mutated features significantly enriched the GlpR regulon (p-value<0.001, permutation test, mutations to feature>1, Figure 3.4a). The GlpR regulon represents the repression of glycerol transport and metabolism in the presence of glucose and the absence of glycerol or glycerol-3-phosphate for *E. coli* K-12 MG1655 [31]. The *glpR* genes of some *E. coli* K-12 MG1655 are pseudogenized. Similar to the U00096.2 *E. coli* K-12 MG1655 reference genome [32], the starting strain used in this experiment has a functional *glpR* [27]. The *glpFKX* operon mutations also significantly converged on the glycerolipid metabolism pathway (Figure B.1), which describes the chemical reactions involving any lipid with a glycerol backbone. The mutations to *glpK* may be working to increase the reaction rate of its product, glycerol kinase (GlpK), which catalyzes the phosphorylation of glycerol and is a rate-limiting enzyme in glycerol metabolism. GlpK is allosterically inhibited by fructose-1,6-bisphosphate (FBP) [33] and the *crr* gene product, Enzyme II A (IIAGlc) [34]. Two SNPs were found in an FBP binding site of GlpK, and 1 SNP substituted an amino acid that had been previously observed to abolish FBP regulation altogether (Table B.2). Ten SNPs targeted regions that are used in forming the GlpK oligomer and two SNPs were predicted to affect structural stability.

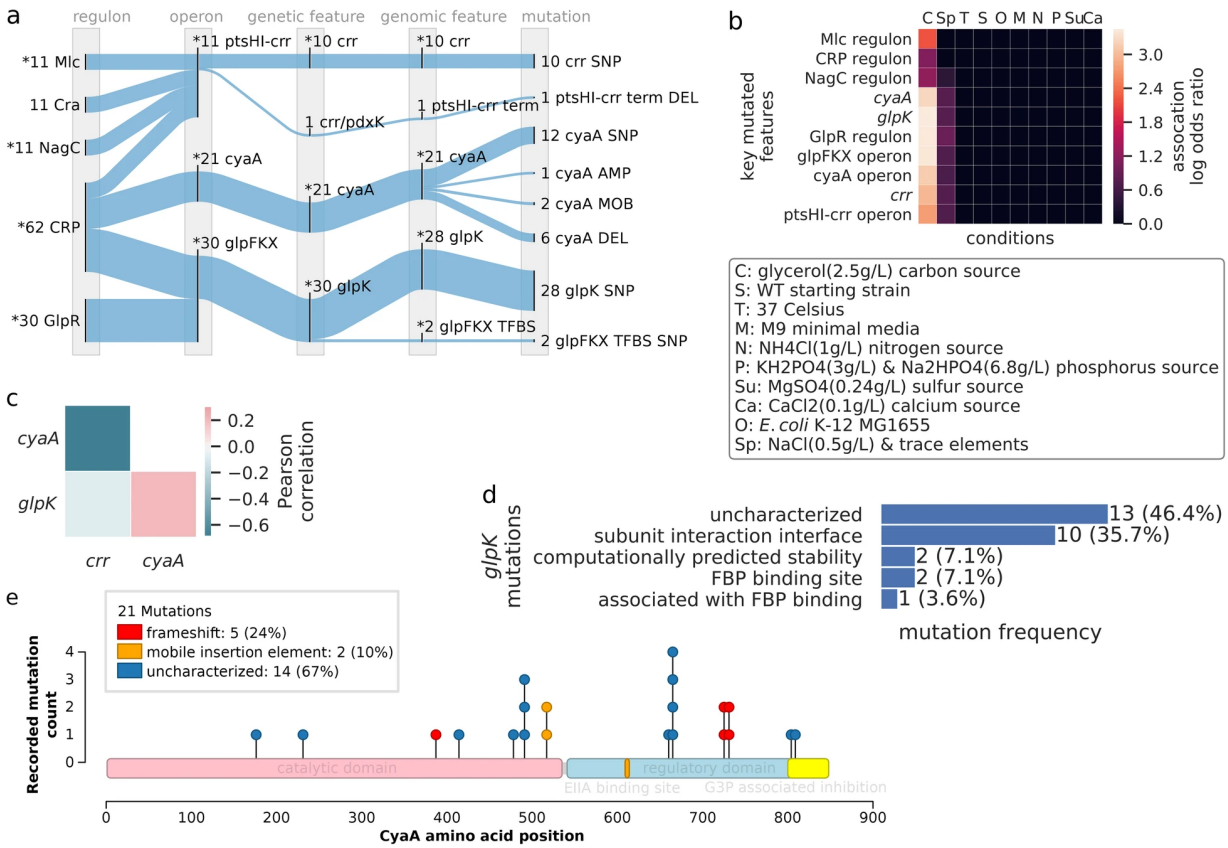


Figure 3.4: The relationships between mutated features regulated by CRP, their associations to glycerol as a carbon source, and the mutation effects on their coding sequences. (a) Mutated features that concurrently converge onto the CRP, Mlc, and GlpR regulons; an asterisk (*) denotes those features found to be mutated with a statistically significant amount relative to all mutations within the ALE experiment (textitp-value<0.05, permutation test). The *glpK*, *crr*, and *cyaA* genes of interest all converge onto the CRP regulon. (b) A heatmap of the log odds ratios for features from Figure 3.4a involved in significant mutation convergence (textitp-value<0.05, permutation test, mutations to feature>1) and their significantly associated conditions (textitp-value<0.01, Fisher's exact test). "att" is used to abbreviate "attenuator terminator". (c) Heatmap of the Pearson correlation of mutation co-occurrence between *cyaA* and *crr*. (d) The frequency of characterizations to *glpK* mutations. e Mutation needle plot demonstrating the types of mutations in the amino acid sequence of the gene *cyaA*. The large light-blue domain is the regulatory domain, the small orange domain is the EIIAGlc binding site, and the yellow domain is the G3P associated inhibition region

GlpK is found to either form a tetramer or dimer, where FBP can inhibit the tetramer's catalytic reaction. Mutations to the subunit interface regions that bias GlpK towards dimer formation, and therefore avoid FBP inhibition, have been seen in a similar glycerol evolution study [35].

The 2 SNPs that were predicted to affect structural stability may be accomplishing the same result. Overall, 53% of the mutations to *glpK* have effects that could disable inhibition by FBP (Figure 3.4d).

Though there were approximately twice as many mutations in *cyaA* as there were to *crr*, these mutations may have accomplished similar adaptive effects. The presence of mutations to either of these targets was inversely correlated with the other (Figure 3.4c), indicating antagonistic epistasis. Adenylate cyclase (AC), *cyaA*'s product, synthesizes cyclic adenosine monophosphate (cAMP). A cAMP molecule will bind with CRP and promote the transcription of CCR genes involved in secondary carbon source metabolism, including the *glpFKX* operon. Before cAMP can do so, this functionality of AC must be activated by binding with a phosphorylated IIAGlc. In the presence of phosphotransferase system (PTS) sugars such as glucose, IIAGlc will instead bind and inhibit GlpK. In the absence of PTS sugars, IIAGlc will become phosphorylated, bind, and activate AC's ability to produce cAMP, ultimately reducing the effect of CCR [36].

A large proportion of the mutations to the *cyaA* gene had a disruptive effect. *cyaA* hosted a total of 21 unique mutations. The mutations that affect *cyaA* were of different types: SNPs, deletions, and mobile element insertions (MOB). One-third of the mutations to *cyaA*, the two MOBs and five frameshifting deletions, had disruptive effects on the coding sequence (Figure 3.4e). AC is thought to have an N-terminal catalytic domain and C-terminal regulatory domain [37]. IIAGlc and glycerol-3-phosphate (G3P) are both associated with interactions on the regulatory domain, where phosphorylated IIAGlc is known to bind to amino acid 609 and activate cAMP production [36], and G3P is thought to lower AC's cAMP production through a feature within the C-terminal's final 48 amino acids [38]. The disruptive mutations to cyaA each affected the subset of AC's features downstream of the mutation (Figure 3.4e). The variety

of features affected, including the features necessary for AC's activity, provides evidence of the non-essentiality of AC in this evolution.

The evidence of potential mutation effects to *crr* also suggested the possible non-essentiality of IIAGlc in this evolution. The convergence of mutations to *crr* resulted in statistically significant enrichment of the Mlc and NagC regulons (Figure 3.4a, Mlc p-value<0.001, NagC P $= 0.0007$, permutation test, mutations to feature>1), which both describe the regulation of the PTS system, a key contributor to the CCR system. All mutations to *crr* landed in the PTS EIIA type-1 domain, which hosts the IIAGlc phosphorylation site. 70% of the mutations to *crr* were predicted to have a structurally destabilizing change on their host structure ($\Delta\Delta$Gpred>2, *crr* Table B.1). 90% of mutations to *crr* were predicted to have deleterious consequences to conserved regions (SIFT score>0.05). Additional evidence towards the mutations having a disruptive effect on *crr* is the possible epistatic relationship between *cyaA* and *crr* mutations along with the mutations to *cyaA* having the clear potential to disrupt its functionality. This evidence suggested the potential non-essentiality of IIAGlc in this evolution.

This ALE experiment generated 148 mutations, an amount prohibitive to comprehensive manual functional analysis. The methods of top-down functional analysis and selection pressure associations generated valuable evidence that revealed a potential solution for increasing glycerol metabolism while maintaining the repression of metabolic systems for other secondary carbon sources. Of the many mutated regulons for this ALE experiment, three were emphasized by being both significantly converged upon by mutations (all p-value<0.001, permutation test, mutations to feature>1) and significantly associated with the designed selection pressure of growth on glycerol (all p-value<0.001, Fisher's exact test). The functions of these regulons contextualized their mutated genes, further informing on the biological systems potentially targeted by adaptation.

Further investigation into the mutated genes informed on the possible beneficial mechanistic effects of their mutations. The mutations to glpK suggest the increase in its reaction rate through the disruption of an inhibition mechanism. The mutations to *cyaA* and *crr* suggest the disruption of cAMP synthesis, resulting in CCR maintenance in the presence of a carbon source that would normally dampen CCR. Such induction of CCR with mutations resulting from an experimental evolution on glycerol has been previously observed [35]. These results serve to promote the value of the evidence generated by the methods of this work in finding and understanding key mutations for ALE experiments with many mutations. There remains more mutations in this ALE experiment, though the mutations to *glpK*, *cyaA*, and *crr* had the strongest and most interpretable signals of adaptation.

Top-down functional analysis and mutated feature-to-condition associations for an *E. coli* K-12 MG1655 derived strain and growth selection for L-serine tolerance highlights changes to the glycine cleavage and transport system and global regulators

The methods of top-down functional analysis and mutated feature-to-condition association analysis contextualized mutations in an ALE experiment tolerizing a genetically engineered *E. coli* K-12 MG1655 strain to L-serine [23], resulting in previously unreported key mutations. This tolerization ALE experiment involved three independent replicate evolution experiments on a strain of *E. coli* K-12 MG1655 that had been genetically engineered to remove internal L-serine degradation pathways with the intention of higher L-serine production for industrial applications. ALE was necessary to tolerize the strain to the toxicity of high L-serine concentrations. The replicate endpoints were each represented by 2 clonal isolates. Comprehensive whole genome resequencing for the study was performed for this work. Mutations that overlapped

between endpoint samples were only considered once (see Methods). Collectively, 27 mutations unique to endpoints were analyzed for this ALE study. The original Mundhada *et al.* study [23] revealed mutations to *thrA*, *lrp*, *rho*, *argP*, *pykF*, and *eno* contributed to L-serine tolerance and fitness in minimal-media.

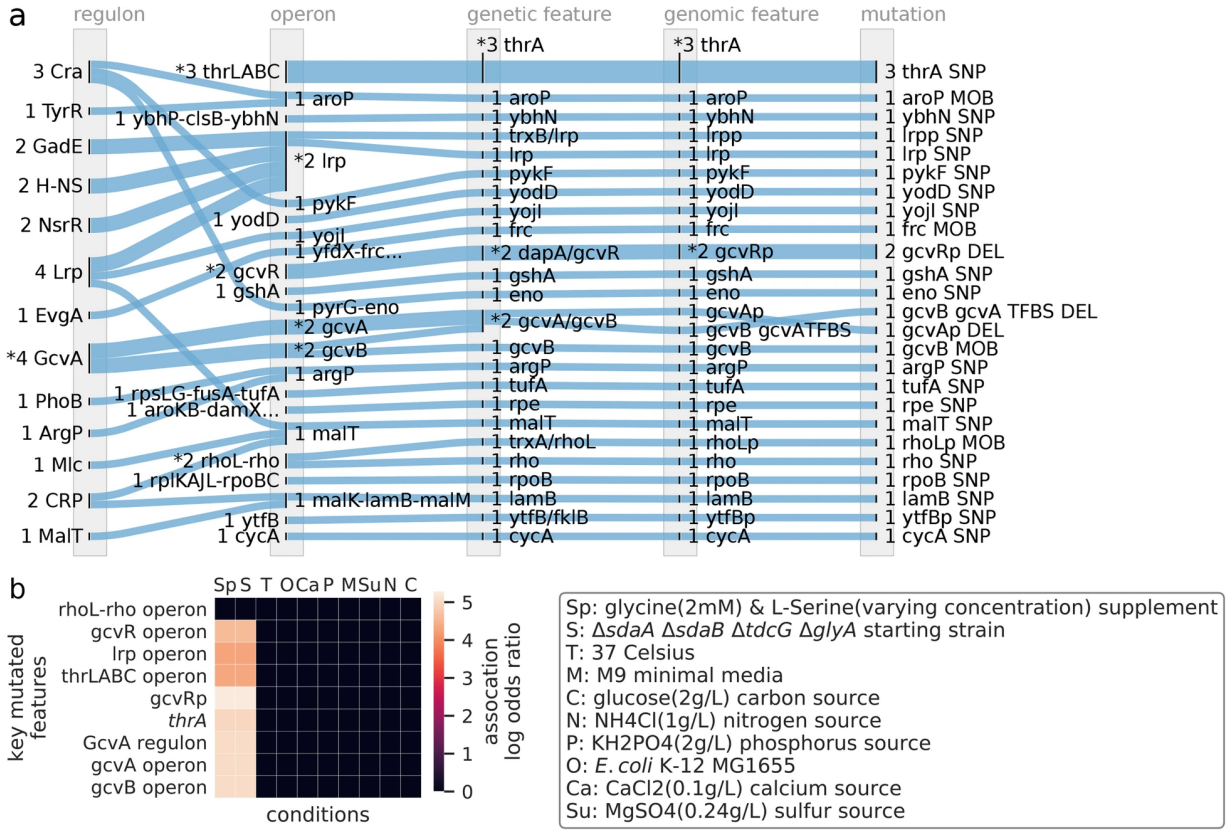


Figure 3.5: The amount of identified key mutations increase through the bottom-up convergence of mutated features on a multiscaled annotation framework. (a) The convergence of mutated genes and regulatory features onto pntAB and sthA operons from a Δ*pgi* ALE experiment on *E. coli* K-12 MG1655 [21, 25]. (b) Proportions of mutated features according to different genomic feature types. The distributions were generated by finding the proportion of each mutated feature per ALE experiment. (c) The proportions of mutated feature types across ALE experiments that were involved in significant convergence (textitp-value<0.05, permutation test, mutations to feature>1) on the multiscale annotation framework. The number of times each feature type was observed mutated is included ('n'). (d) Proportion of mutated features across ALE experiments that had significant convergence up to an annotation type on the multiscale annotation framework of Figure 3.1

The GcvA regulon, representing the glycine cleavage function, was significantly converged upon (p-value = 0.0013, permutation test, mutations to feature>1) by mutated features (Figure 3.5a). The starting strain for this ALE experiment was auxotrophic for glycine, therefore glycine was added to the media. Mutations that could offset a glycine deficiency would be beneficial. The GcvA regulon was also found to be significantly associated with the starting strain mutations and L-serine tolerance (Figure 3.5b).

The operon level of annotations contained the largest amount of significantly convergent features for this ALE experiment, where all significantly mutated operons (gvcA p-value = 0.0046, remainder p-value<0.001, permutation test, mutations to feature>1) except *rhoL-rho* were significantly associated (p-value<0.001, Fisher's exact test) with L-serine tolerance and starting strain mutations (Figure 3.5b). Due to hosting the largest amount of significantly convergent features, the operon level of annotations was the most revealing in the adaptive changes for this ALE experiment. By observing the mutations to operons, we can better understand non-coding mutations and their potential effect on gene products and the systems they contribute to. Regulatory genes for the glycine cleavage system host mutations in the regulatory features of their transcription units. The *gcvA* operon contains mutations in its promoter and a repressive GcvA TFBS (Figure 3.5a). The *gcvA* gene encodes for a transcriptional dual regulator for the glycine cleavage system operon *gcvTHP* [39, 40]. Downregulating *gcvA*'s transcription could inhibit its activation of *gcvTHP*'s transcription. GcvA can also form a glycine cleavage repression complex with GcvR [40]. The *gcvR* gene's promoter (gcvRp) experienced a 1 bp DEL in two endpoints (Figure 3.5a). Upregulating *gcvR*'s transcription could have the effect of further repressing the gcvTHP operon coding for glycine cleavage. GcvR is additionally inhibited by glycine [41]. Mutations to *gcvA* and *gcvR* regulatory features may simply be removing their presence through the

alteration of their promoters, leaving the glycine cleavage system operon with an unstimulated transcription rate. The *gcvB* operon, which plays a role in the glycine transport system and is co-regulated with the glycine cleavage system through GcvA, hosted mutations to two different features. The *gcvB* gene of one endpoint hosted a mobile element insertion (i.e., MOB), where another endpoint hosted a 1 bp DEL in an activating GcvA TFBS (Figure 3.5a). The *gcvB* gene encodes a small regulatory RNA that acts as a repressor of *cycA* [42], which also hosts a SNP, and functions as a symporter of glycine, D-serine, and D-alanine [43]. Disruption of *gcvB* may increase the uptake of glycine by disabling the repression of CycA. The original study recognized the mutations to the *gcvB* operon as potentially beneficial. The emphasis on the glycine cleavage system due to the significant convergence of the GcvA regulon by mutations suggests mutations to the *gcvA* and *gcvR* operons as additionally being key.

The operons for global regulators *lrp* and *rho* were both significantly enriched by the convergence of their mutated coding sequences and promoters (p-value<0.001, permutation test, mutations to feature>1). *lrp* is a global regulator for the leucine response system and has experimental evidence demonstrating its participation in L-serine tolerance [23]. The roles of these global regulators in this adaptation remains unknown due to their large network of interactions, though the *lrp* and *rho* coding sequence mutations were found to contribute the most fitness among all endpoint mutations of this experiment [23].

This ALE experiment generated a variety of mutations to both coding and non-coding regions, where mutations to non-coding features not considered in the original study provided evidence for new key mutations. The methods of top-down functional analysis and selection pressure associations generated valuable evidence that revealed a potential optimization for the strain's glycine auxotrophy. The richer annotations for non-coding features on the genome revealed the

high frequency of regulatory features targeted by ALE mutations and facilitated the discovery of previously unrecognized mutated regulatory features. Significant enrichment of features on the multiscale annotation framework emphasized the systems in which the ALE mutations affected, informing on the normally more elusive non-coding mutations. This emphasis on higher-level features revealed previously unconsidered changes to the glycine cleavage system that could benefit the host strain. These results serve to promote the value of the evidence generated by the methods of this work in finding and understanding key mutations.

3.4 Discussion

To better identify and understand causal mutations from ALE experiments, this study developed a multiscale genome annotation framework and applied a statistical enrichment method for mutated features across annotation scales. Using this framework, it was found that (1) ALE mutations target a variety of regulatory features including promoters, attenuator-terminators, and ribosomal binding sites, (2) mutated non-coding regulatory features were often involved in significant convergence events, and (3) the method of bottom-up convergence from small to large features on the multiscale annotation framework found more key mutations than when considering only genes and intergenic regions for mutation convergence. The convergence of mutated features onto broader functional annotations additionally enabled a top-down approach to mutation functional analysis, where one can first consider biological functions hosting mutations before investigating the numerous smaller mutated features. Further, we computed statistical associations between a large set of experimental evolution conditions and mutated features. These associations provided evidence towards clarifying the selection pressures for mutated features among the multiple conditions describing each ALE experiment. Taken together, the results

presented here have several implications.

First, the identification of key mutations was enhanced by seeking mutations involved in statistically significant convergence events across multiple scales of genome annotation. By including regulatory feature annotations alongside genes, a larger proportion of an experiment's mutations were placed in features with known functions, which were subsequently connected to broader annotation types such as operons and functional annotations. Being able to connect more mutations to known features resulted in more convergence events across the annotation framework. Additionally, the application of a statistical enrichment test can be used to prioritize convergence events in the followup work of mutation functional analysis. This will become more important as ALE experiments increase in scale. All mutated regulatory feature types were involved in statistically significant convergence events, and except for promoters, had a higher median proportion of their mutations involved in these events than mutated genes. These results suggest that mutations to non-coding regulatory features should be considered as important as coding features when searching for key mutations, though standard genome annotations often lack non-coding regulatory features. When only considering mutations from significantly converged upon features, the amount of significant or key mutations on larger-scale annotation types (operons, pathways, regulons, COGs) remains larger than those on smaller annotation types (e.g., genes, promoters). The convergence of a variety of small mutated features onto broader annotations also serves to maximize the amount of evidence used in identifying the overall changes of an adaptive genotype. In its application, this method identified new regulatory key mutations for a published ALE study on L-serine tolerance [23], demonstrating the value of richer annotations and convergence across multiple scales of annotation. Additionally, significant convergence involving the most beneficial mutations to the L-serine tolerance evolution, those to the coding

sequences and regulatory features of the *lrp* and *rho* genes, was made explicit with the inclusion of regulatory features in mutation annotations. These approaches can therefore lessen the challenge of finding key mutations by providing enhancements to traditional methods.

Second, the functional analysis of key mutations was enhanced by considering the functions of the features upon which mutations converged and the conditions that mutated features were associated with. We found that a top-down mutation functional analysis proved valuable in contextualizing key mutations according to the high-level functional annotations to which they were connected and shared with other key mutations. The three most frequently mutated genes of strains from the ALE experiment on glycerol [22] all converged on the CRP regulon, revealing the functional proximity of their products through their participation in the CCR system. This convergence served to inform how mutations to these genes could have manipulated the CCR system to enable only the catabolism of glycerol and maintain the repression of the remaining secondary carbon source catabolism systems. The new key mutations proposed for the L-serine ALE study [23] were additionally enhanced by this top-down functional analysis approach with the significant convergence of the GcvA regulon, describing their involvement in the glycine cleavage system. The starting strain for this experimental evolution was auxotrophic for glycine and had to have it provided as a supplement. This convergence informed how these mutated features could alleviate the strain's glycine auxotrophy by either repressing the glycine cleavage system to conserve glycine or increasing the uptake of glycine. Similarly, the association of mutated features to conditions can deconvolute the conditions selecting for mutations. The three most prominent key mutated genomic features of the glycerol ALE were primarily associated with the condition of glycerol as a carbon source, further strengthening the functional analysis' focus on glycerol metabolism. The key mutations of the L-serine ALE with genomic features mutated

more than once were likewise uniquely associated with the conditions of the starting strain and high concentrations of L-serine. Mutated feature-to-condition association is valuable to mutation interpretation in that researchers can focus their efforts on how these mutations enable adaptation to the significantly associated conditions of interest. These methods can lessen the challenge of understanding key mutations by providing enhancements to traditional methods.

The methods presented here demonstrate the value in leveraging the diverse available resources describing ALE variants, but are not without limitations. The procedure for finding the statistical significance of mutation convergence assumes that each nucleotide on the genome has the same probability of being spontaneously mutated, leading to features having the probability of being mutated proportional to their length in nucleotides. Nucleotides can have varying mutation rates; for example, wild type E. coli indel mutation rates have been observed to be higher in mononucleotide runs of 4 or more [18]. Parameterizing different genomic features or nucleotide locations with better representative mutation rates would enable increased accuracy in the significance measurements of mutation convergence events. There additionally exist numerous annotation frameworks currently not integrated, such as structural annotations [44, 45], gene product complexes [46], and gene ontologies [47, 48]. Their inclusion would increase the coverage of biological domains for the methods this work. The abundance of annotation frameworks is often limited to model organisms such as *E. coli* K-12 MG1655. To use these methods with other organisms would require knowledge bases that describe the genes and regulatory architecture of those organisms' genome, along with the biological functions that the genes contribute to. Evolution experiments often also include midpoint or intermediate samples. The results of this work only included endpoint samples, as not all experiments had midpoints, to enable a uniform analysis method across all experiments. The inclusion of midpoint samples could provide further

evidence of mutation convergence through the dynamics of clonal interference, as well as the opportunity for mutation time series analysis. Correlations between mutated features proved useful for interpreting the relationship between the mutated *cyaA* and *crr* genes from the glycerol carbon source ALE case study. Applying the same correlation method on a larger scale mutated feature set tends to generate less interpretable results, though statistical evidence of relationships between mutated features would still be valuable. More sophisticated methods for establishing these relationships within large-scale sets of mutated features are necessary to extract meaningful interpretations. Finally, the key mutations found in this work were derived from sets of mutations of a scale that still can be managed manually, albeit with great effort. As ALE datasets grow and automation-enabled studies increase in overall sample number, this work's methods will become necessary to comprehensively consider and understand an ALE experiment's mutations.

3.5 Conclusion

Here, we have reported on data-driven approaches to better find and understand ALE key mutations. We demonstrated how a multiscale genome annotation framework and statistical association methods can better identify and characterize adaptive mutations generated through controlled evolution experiments. We anticipate that this workflow will be leveraged in the future to provide deeper insights into the vast amounts of -omics data that are generated for targeted microorganisms amenable to ALE and provide the blueprints for similar data-driven annotation frameworks.

Acknowledgements

The authors would like to thank Laurence Yang, Anand Sastry, Jean-Christophe Lachance, Ke Chen, Amitesh Anand, Bin Du, and Colton Llyod for their helpful discussions. The authors would like to thank Josefin Johnsen and Marc Abrams for their editorial comments.

Conceptualization: PVP, AMF, BOP. Technical support: MW. Conceptual support: JTY, DH, ZAK, JT, TES. Writing: PVP, JTY, DH, AMF.

This work was funded by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (grant number NNF10CC1016517). Funding was also provided by the NIH National Institute of Allergy and Infectious Diseases (grant number U01AI124316). JTY was funded by the Institute for Systems Biology's Translational Research Fellows Program. The funding agencies had no role in the study design, data collection, analysis and interpretation, decision to publish, or preparation of the manuscript.

Chapter 3 is a reprint of material published in: **Patrick V. Phaneuf**, Patrick V. Phaneuf, James T. Yurkovich, David Heckmann, Muyao Wu, Troy E. Sandberg, Zachary A. King, Justin Tan, Bernhard O. Palsson, Adam M. Feist "Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity". BMC Genomics. 2020 Jul 25;21(1):514. The dissertation author is the primary author.

3.6 References

1. Notley-McRobb, L. & Ferenci, T. The generation of multiple co-existing mal-regulatory mutations through polygenic evolution in glucose-limited populations of Escherichia coli. en. *Environ. Microbiol.* **1,** 45–52. ISSN: 1462-2912 (Feb. 1999).

2. Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. & Botstein, D. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. en. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16144–16149. ISSN: 0027-8424 (Dec. 2002).

3. Araya, C. L., Payen, C., Dunham, M. J. & Fields, S. Whole-genome sequencing of a laboratory-evolved yeast strain. en. *BMC Genomics* **11,** 88. ISSN: 1471-2164 (Feb. 2010).

4. Yona, A. H., Manor, Y. S., Herbst, R. H., Romano, G. H., Mitchell, A., Kupiec, M., Pilpel, Y. & Dahan, O. Chromosomal duplication is a transient evolutionary solution to stress. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 21010–21015. ISSN: 0027-8424, 1091-6490 (Dec. 2012).

5. Chang, S.-L., Lai, H.-Y., Tung, S.-Y. & Leu, J.-Y. Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. en. *PLoS Genet.* **9,** e1003232. ISSN: 1553-7390, 1553-7404 (Jan. 2013).

6. Sandberg, T. E., Salazar, M. J., Weng, L. L., Palsson, B. O. & Feist, A. M. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. en. *Metab. Eng.* ISSN: 1096-7176, 1096-7184 (Aug. 2019).

7. Van den Bergh, B., Swings, T., Fauvart, M. & Michiels, J. Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. en. *Microbiol. Mol. Biol. Rev.* **82.** ISSN: 1092-2172, 1098-5557 (Sept. 2018).

8. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Appl. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

9. Barrick, J. E. & Lenski, R. E. Genome-wide mutational diversity in an evolving population of Escherichia coli. en. *Cold Spring Harb. Symp. Quant. Biol.* **74,** 119–129 (Sept. 2009).

10. Gresham, D. & Dunham, M. J. The enduring utility of continuous culturing in experimental evolution. en. *Genomics* **104,** 399–405. ISSN: 0888-7543, 1089-8646 (Dec. 2014).

11. Tenaillon, O., Rodriguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D. & Gaut, B. S. The molecular diversity of adaptive convergence. en. *Science* **335,** 457–461. ISSN: 0036-8075, 1095-9203 (Jan. 2012).

12. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. en. *Nucleic Acids Res.* **37,** 1–13. ISSN: 0305-1048, 1362-4962 (Jan. 2009).

13. Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., Alquicira-Hernandez, K., Martinez-Flores, I., Pannier, L., Castro-Mondragon, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martinez, C., Perez-Rueda, E., Alquicira-Hernandez, S., Porron-Sotelo, L., Lopez-Fuentes, A., Hernandez-Koutoucheva, A., Del Moral-Chavez, V., Rinaldi, F. & Collado-Vides, J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. en. *Nucleic Acids Res.* **44,** D133–43. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

14. Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., Gerdes, S., Henry, C. S., Kenyon, R. W., Machi, D., Mao, C., Nordberg, E. K., Olsen, G. J., Murphy-Olson, D. E., Olson, R., Overbeek,

R., Parrello, B., Pusch, G. D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H. & Stevens, R. L. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. en. *Nucleic Acids Res.* **45,** D535–D542. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

15. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. en. *Nucleic Acids Res.* **28,** 33–36. ISSN: 0305-1048 (Jan. 2000).

16. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* **47,** D1164–D1171. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

17. Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sanchez-Perez, M., Gomez-Romero, L., Ledezma-Tejeida, D., Garcia-Sotelo, J. S., Alquicira-Hernandez, K., Muniz-Rascado, L. J., Pena-Loredo, P., Ishida-Gutierrez, C., Velazquez-Ramirez, D. A., Del Moral-Chavez, V., Bonavides-Martinez, C., Mendez-Cruz, C.-F., Galagan, J. & Collado-Vides, J. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. en. *Nucleic Acids Res.* **47,** D212–D220. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

18. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2774–83. ISSN: 0027-8424, 1091-6490 (Oct. 2012).

19. Miller, M. L., Reznik, E., Gauthier, N. P., Aksoy, B. A., Korkut, A., Gao, J., Ciriello, G., Schultz, N. & Sander, C. Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. en. *Cell Syst* **1,** 197–209. ISSN: 2405-4712 (Sept. 2015).

20. Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Medigue, C., Schneider, D. & Lenski, R. E. Tempo and mode of genome evolution in a 50,000-generation experiment. en. *Nature* **536,** 165–170 (Aug. 2016).

21. Long, C. P., Gonzalez, J. E., Feist, A. M., Palsson, B. O. & Antoniewicz, M. R. Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 222–227 (Jan. 2018).

22. LaCroix, R. A., Palsson, B. O. & Feist, A. M. A Model for Designing Adaptive Laboratory Evolution Experiments. en. *Appl. Environ. Microbiol.* **83** (Apr. 2017).

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

24. Conrad, T. M., Joyce, A. R., Applebee, M. K., Barrett, C. L., Xie, B., Gao, Y. & Palsson, B. O. Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term

laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. en. *Genome Biol.* **10,** R118. ISSN: 1465-6906 (Oct. 2009).

25. Charusanti, P., Conrad, T. M., Knight, E. M., Venkataraman, K., Fong, N. L., Xie, B., Gao, Y. & Palsson, B. O. Genetic Basis of Growth Adaptation of Escherichia coli after Deletion of pgi, a Major Metabolic Gene. *PLoS Genet.* **6,** e1001186 (Nov. 2010).

26. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System. en. *Metab. Eng.* **48,** 233–242. ISSN: 1096-7176, 1096-7184 (July 2018).

27. Phaneuf, P. *SBRG/bop27refseq: Zenodo DOI release (Version v1.4.1)* Dec. 2019.

28. Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R. & Palsson, B. O. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. en. *Nat. Genet.* **38,** 1406–1412. ISSN: 1061-4036 (Dec. 2006).

29. Deutscher, J. The mechanisms of carbon catabolite repression in bacteria. en. *Curr. Opin. Microbiol.* **11,** 87–93. ISSN: 1369-5274 (Apr. 2008).

30. Gorke, B. & Stulke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. en. *Nat. Rev. Microbiol.* **6,** 613–624. ISSN: 1740-1526, 1740-1534 (Aug. 2008).

31. Larson, T. J., Cantwell, J. S. & van Loo-Bhattacharya, A. T. Interaction at a distance between multiple operators controls the adjacent, divergently transcribed glpTQ-glpACB operons of Escherichia coli K-12. en. *J. Biol. Chem.* **267,** 6114–6121. ISSN: 0021-9258 (Mar. 1992).

32. Freddolino, P. L., Amini, S. & Tavazoie, S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. en. *J. Bacteriol.* **194,** 303–306. ISSN: 0021-9193, 1098-5530 (Jan. 2012).

33. Zwaig, N. & Lin, E. C. Feedback inhibition of glycerol kinase, a catabolic enzyme in Escherichia coli. en. *Science* **153,** 755–757. ISSN: 0036-8075 (Aug. 1966).

34. Novotny, M. J., Frederickson, W. L., Waygood, E. B. & Saier Jr, M. H. Allosteric regulation of glycerol kinase by enzyme IIIglc of the phosphotransferase system in Escherichia coli and Salmonella typhimurium. en. *J. Bacteriol.* **162,** 810–816. ISSN: 0021-9193 (May 1985).

35. Applebee, M. K., Joyce, A. R., Conrad, T. M., Pettigrew, D. W. & Palsson, B. O. Functional and metabolic effects of adaptive glycerol kinase (GLPK) mutants in Escherichia coli. en. *J. Biol. Chem.* **286,** 23150–23159. ISSN: 0021-9258, 1083-351X (July 2011).

36. Saier Jr, M. H., Kukita, C. & Zhang, Z. Transposon-mediated directed mutation in bacteria and eukaryotes. en. *Front. Biosci.* **22,** 1458–1468. ISSN: 1945-0494, 1093-4715 (Mar. 2017).

37. Roy, A., Danchin, A., Joseph, E. & Ullmann, A. Two functional domains in adenylate cyclase of Escherichia coli. en. *J. Mol. Biol.* **165,** 197–202. ISSN: 0022-2836 (Mar. 1983).

38. Eppler, T., Postma, P., Schutz, A., Volker, U. & Boos, W. Glycerol-3-phosphate-induced catabolite repression in Escherichia coli. en. *J. Bacteriol.* **184,** 3044–3052. ISSN: 0021-9193 (June 2002).

39. Stauffer, L. T. & Stauffer, G. V. GcvA interacts with both the alpha and sigma subunits of RNA polymerase to activate the Escherichia coli gcvB gene and the gcvTHP operon. en. *FEMS Microbiol. Lett.* **242,** 333–338. ISSN: 0378-1097 (Jan. 2005).

40. Ghrist, A. C., Heil, G. & Stauffer, G. V. GcvR interacts with GcvA to inhibit activation of the Escherichia coli glycine cleavage operon. en. *Microbiology* **147,** 2215–2221. ISSN: 0026-2617, 1350-0872 (Aug. 2001).

41. Heil, G., Stauffer, L. T. & Stauffer, G. V. Glycine binds the transcriptional accessory protein GcvR to disrupt a GcvA/GcvR interaction and allow GcvA-mediated activation of the Escherichia coli gcvTHP operon. en. *Microbiology* **148,** 2203–2214. ISSN: 0026-2617, 1350-0872 (July 2002).

42. Pulvermacher, S. C., Stauffer, L. T. & Stauffer, G. V. Role of the sRNA GcvB in regulation of cycA in Escherichia coli. en. *Microbiology* **155,** 106–114. ISSN: 0026-2617, 1350-0872 (Jan. 2009).

43. Ghrist, A. C. & Stauffer, G. V. The Escherichia coli glycine transport system and its role in the regulation of the glycine cleavage enzyme system. en. *Microbiology* **141 (Pt 1),** 133–140. ISSN: 0026-2617, 1350-0872 (Jan. 1995).

44. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. en. *Nucleic Acids Res.* **28,** 235–242. ISSN: 0305-1048 (Jan. 2000).

45. UniProt Consortium, T. UniProt: the universal protein knowledgebase. en. *Nucleic Acids Res.* **46,** 2699. ISSN: 0305-1048, 1362-4962 (Mar. 2018).

46. Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I. & Karp, P. D. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. en. *Nucleic Acids Res.* **45,** D543–D550. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

47. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. en. *Nat. Genet.* **25,** 25–29. ISSN: 1061-4036 (May 2000).

48. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. en. *Nucleic Acids Res.* **47,** D330–D338. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

Chapter 4

Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data

4.1 Abstract

Microbes are being engineered for an increasingly large and diverse set of applications. However, the designing of microbial genomes remains challenging due to their complexity and existing knowledge gaps within biological systems. Bioengineers have a unique advantage over other engineering disciplines: biological systems have been solving such challenges long before human intervention through adaptive evolution. Adaptive Laboratory Evolution (ALE) methods leverage these natural problem-solving processes to elucidate biological mechanisms and generate application solutions. The promise of ALE methods has resulted in the generation of a substantial

amount of public ALE data, which, in aggregate, can contribute novel insights towards ALE-derived strain design. This work presents a meta-analysis on aggregated mutational data for the biotechnological host Escherichia coli, leveraging rich mutation annotations and multiple mutation characterization methods. Three novel experimentally validated designs relevant to metabolic engineering applications are presented as use cases of such a framework and analysis. These results demonstrate how fundamental principles can be extracted from aggregated ALE data to enhance strain design efforts.

4.2 Introduction

The ability to precisely engineer microbial strains continues to improve. Multiple fields have emerged to take advantage of recent advances in biomolecular methods and technologies [1]. However, an incomplete understanding of biology renders the rational design of microbial strains challenging [2]. Introducing rationally designed changes into the DNA of a cellular chassis often leads to a perturbed suboptimal metabolic or regulatory state, resulting in the underachievement of goals [3] and the need for a prolonged engineering effort that can require up to 6–8 years and over $50 million in costs [2].

Bioengineers have a unique advantage over other engineering disciplines: biological systems have been solving challenges long before human intervention through adaptive evolution. The opportunity thus exists to leverage these built-in problem-solving processes to discover and elucidate biological functions and generate solutions for applications. Adaptive Laboratory Evolution (ALE) is the formalization of a controlled evolutionary process that can successfully be applied to understand and engineer bacterial strains. ALE can provide adaptive mutations to optimize a strain's growth rate (or related fitness value), a foundational property for microbial

research and applications [4]. When a production pathway (e.g., metabolic) can be coupled with growth, ALE can provide adaptive mutations that will increase the throughput of this pathway [4]. Furthermore, if a strain's fitness has been severely disrupted by a designed change, ALE can identify mutations that rebalance the cell's homeostasis [4]. ALE can additionally serve to harden strains against industrial conditions and improve their utilization of secondary or non-native substrates [4].

Due to ALE's potential for discovery and application [4], almost 700 papers and over 18,000 experimental evolutions has been published [**Van˙den˙Bergh2018**]. The growth in ALE data has inspired multiple efforts towards its consolidation and analysis [**Van˙den˙Bergh2018**, 5–7]. Aggregating public data is expected to enable new discoveries not evident through single experiments [8]. Critical biomedical efforts, such as cancer and antimicrobial resistance research, have benefited from the meta-analysis of big data sets describing their respective fields [9–11]. Previous work has datamined public ALE data, though their results were limited to genetic loci and therefore didn't explore the specific mutational sequence changes [5]. Nucleotide-level resolution is ideal for strain design and the aggregation of ALE mutation sequence changes could reveal the specific types of changes adaptive evolution selects across defined genomic features and experimental conditions. These ALE-derived design principles could inform strain design efforts [12]. Thus, one could initiate a design based on an ALE mutation or evolved strain that best represents an apparent mutation trend from aggregated mutational data. One could further strive to understand ALE mutation trends well enough to propose novel sequence changes that would accomplish the same fitness benefit, demonstrating the possibility to extract strain design principles from ALE data.

This study seeks to design strain variants with potential for applications by using rich

mutation annotations and structural biology methods on a consolidated ALE dataset of mutations and their experimental conditions. Aggregated ALE mutational and experimental conditions data was leveraged from ALEdb, a web-based platform reporting on experimental evolution mutations and their conditions [6]. The combination of rich annotations and aggregated ALE data enables mutated systems to be associated with experimental conditions to aid in deconvoluting mutation selection pressures [7]. Structural biology methods and functional annotations helped reveal the protein properties being targeted by ALE mutations. These analyses enable the interpretation of ALE mutation objectives for specific conditions of interest, which were leveraged to design and build novel variants of similar benefit to the ALE mutations. This work's results demonstrate how to leverage existing aggregated ALE data and biological knowledge in microbial strain design and form a basis for similar applications in biomedical or basic discovery efforts.

4.3 Results

4.3.1 A workflow leveraging mutation trends in consolidated ALE data to design variants

A generalized workflow for variant design from ALE data was developed and validated (Figure 4.1). The workflow leverages a set of relationships established through the consolidation of multiple ALE experiment mutations and experimental design metadata. The following describes the workflow's steps:

1. *Select conditions of interest.* This step is often dictated by the desired application and is informed by known biological mechanisms associated with the desired phenotype. The conditions available for investigation come from the experimental conditions within the

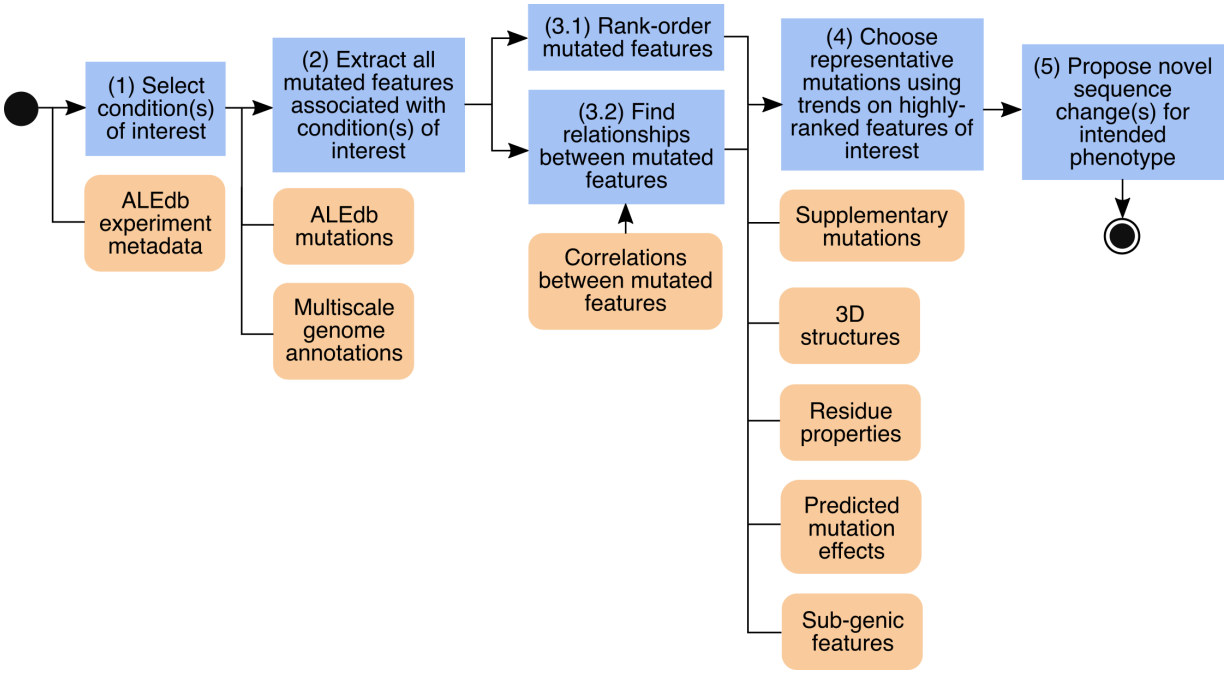


Figure 4.1: A general workflow to derive strain designs from aggregated ALE mutations and experimental conditions.

metadata. As the database and reference set grows, the queryable cases will also increase.

2. *Extract all mutated features associated with conditions of interest.* Querying resources like ALEdb (10) and searching the literature can provide a set of mutated features. Extending mutation annotations beyond the standard types included in genome references can help identify systems beyond genes and intergenic regions that may be associated with conditions of interest [7].

3. *Rank-order the features of interest and identify relationships between them.* Rank-order mutated features of interest by mutation frequency and/or strength of association to condition of interest. Relationships between mutated features should also be considered (e.g., in the case of synergistic or antagonistic epistasis) and can help avoid potential incom-

patible changes and provide insights into the systemic changes that result from multiple mutations.

4. *Choose representative mutations using trends on highly-ranked features of interest.* Identify mutation trends on highly-ranked features of interest and select representative mutations to establish initial ALE variant designs. Trends can be identified through mutation clustering on targets and can be informed using structural biology and mutation effect prediction methods.

5. *Propose novel sequence changes.* Interpret the ALE mutation trend's objectives and propose a novel sequence change that would potentially accomplish the same fitness benefit as the representative ALE mutation. Interpretations can be informed using structural biology and mutation effect prediction methods.

4.3.2 Meta-analysis trends suggest types of designs to expect from ALE data

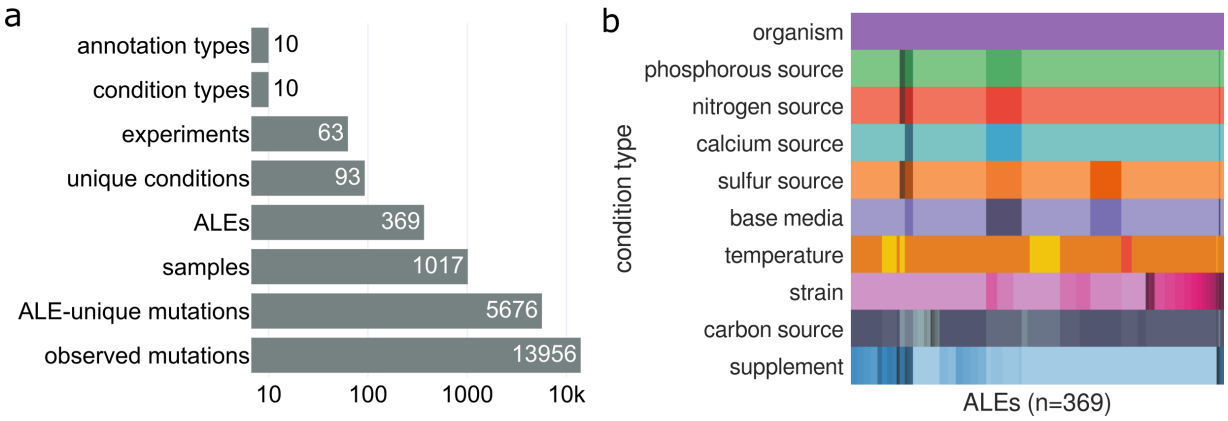


Figure 4.2: (a) The magnitudes for different aspects of the ALE data used within this study. (b) The combination of condition labels across consolidated ALE samples. The mapping between individual colors and labels can be found with the supplementary figures.

This study specifically used ALE mutations and metadata from ALEdb, a database for

experimental evolution mutation data [6], and an enriched set of mutation annotations generated by a multiscale annotation method [7]. The dataset contained 63 *Escherichia coli* K-12 MG1655 based ALE experiments from ALEdb [6], totaling 369 independent evolutions and 5676 ALE-unique mutations (Figure 4.2a). The mutations were annotated with 10 different genome annotation types to enable the identification of mutation convergence on a broad set of genomic features and biological functions [7] (Figure 4.2a, Figure 4.3a). The dataset tracked 10 different condition types describing the strain and environment of ALE experiments with a total of 93 unique conditions (Figure 4.2a). Meta-analysis of this consolidated dataset revealed trends that predict the general shape of the workflow's results and will be described in the following sections.

Non-coding regulatory features are more often disrupted by ALE mutations than coding features

ALE mutations come in a variety of types and can affect a variety of features encoded on the genome. Six different types of mutations were found within the dataset (Figure 4.3a): single nucleotide polymorphisms (SNP), deletions (DEL), insertions (INS), mobile element insertions (MOB), multi-nucleotide substitutions (SUB), and copy number variations (CNV). Different mutation types manifest at different frequencies, resulting in different amounts for each mutation type within this dataset (Figure 4.3a). The mutated feature annotations used in this study range from small genomic features (e.g., terminators) to larger features (i.e., operons, regulons, and pathways) describing biological function (Figure 4.3a). There is a clear trend in the frequency of mutated features: genes are most often mutated, with promoters maintaining very large amounts of mutations relative to the remaining non-coding features. Multi-nucleotide mutations or overlapping features can result in more than one feature affected by a mutation, therefore more

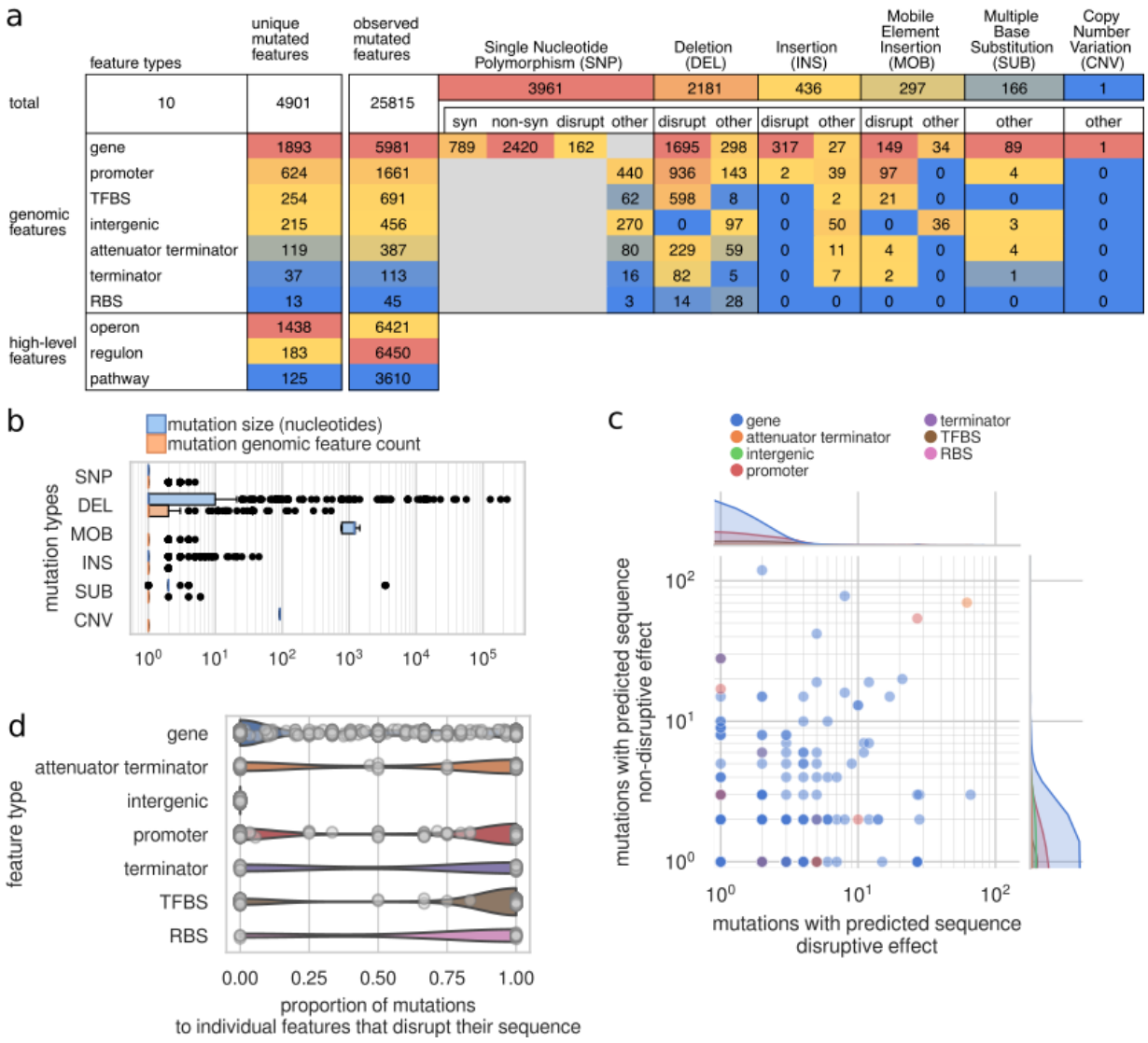


Figure 4.3: (a) Table of mutation type and mutated feature frequencies. Synonymous SNPs are abbreviated as "syn", nonsynonymous SNPs are abbreviated as "non-syn", and mutations disruptive to a feature's sequence are abbreviated as "disrupt". (b) The distribution of mutation sizes and amount of genomic features affected according to mutation types. Abbreviations: SNP, single nucleotide polymorphism; DEL, deletion; MOB, mobile insertion elements; INS, insertion; SUB, substitution; AMP, amplification. (c) The number of sequence disruptive and and non-disruptive mutations for individual genomic features. Abbreviations: TFBS, transcription factor binding site; RBS, ribosomal binding site. (d) The proportion of mutations to individual features across feature types that are sequence disruptive.

mutated features than mutations can occur within an ALE experiment [7] (Figure 4.3a, Figure 4.3b).

Amino acid substitutions and some of the disruptive effects mutations can have on genomic feature sequences can be confidently predicted. These predictions rely on mutation size (Figure 4.3b) or specific coding sequence changes. The effect of mutations on the translation of genes are the most straightforward to predict. SNPs can result in synonymous or nonsynonymous codon changes, where a nonsynonymous SNP can result in an amino acid substitution or a truncation. The effects of small structural variants (SV) are more difficult to predict, though it is likely safe to assume that SVs of 10 nucleotides or more are likely disruptive to a feature's function. It is expected that both SNPs and SVs can disrupt the encoded function of coding features (i.e., premature stop codons, start codon removal, frameshift) and non-coding feature sequences. For the sake of simplicity with the meta-analysis, these types of mutation effects are referred to as "disruptions" (Figure 4.3). SNPs are also the most frequent mutation type and the largest contributor to mutated features (Figure 4.3a). They are most frequently selected for within genes and are most often predicted to result in a non-synonymous amino acid substitution. Deletions are the most frequent SV and contribute a substantial amount of the disruptions to features.

Different feature types display different mutation type and effect trends. All non-coding feature types, besides intergenic regions, are more often targeted by disruptive mutations while genes are more often targeted by non-truncating mutations (Figure 4.3d). This suggests that non-coding features are more likely to be disrupted than refined by ALE mutations. Most features are often mutated with both disruptive and non-disruptive mutations, though there are features more often affected by one of the predicted mutation effects (Figure 4.3c). The features affected by only non-disruptive mutations may be benefiting from gain-of-function mutations, which represent potential for variant designs involving something other than truncations.

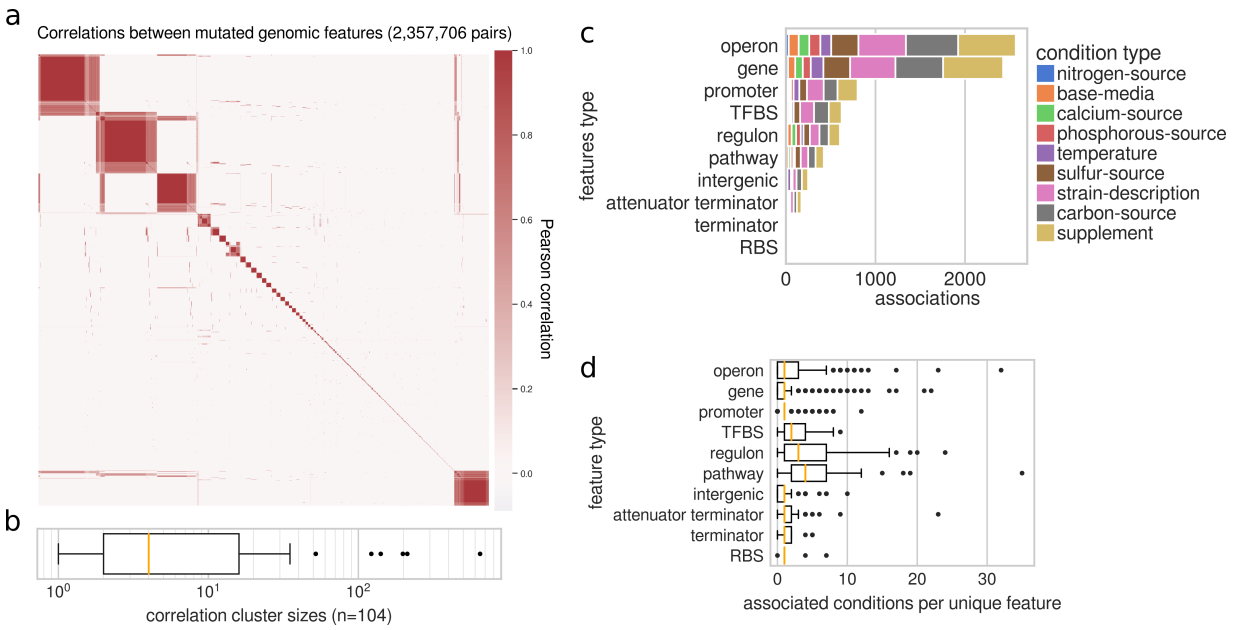


Figure 4.4: (a) A clustermap of the Pearson correlation coefficients for all genomic feature pairs. (b) The distribution of cluster sizes from the clustering of all genomic feature pairs according to their correlation. (c) The total amount of statistically significant associations between unique features and conditions according to feature types. (d) The amount of significantly associated conditions per unique genomic feature.

Most conditions may have a small amount of compatible sequence changes for strain optimization

Correlations between all mutated genomic features (gene, promoter, TFBS, intergenic, attenuator terminator, terminator, RBS) across the entire set of mutations may approximate coarse-grain relationships. Positively correlated genomic features should represent features that can be mutated together to optimize a strain for one or more conditions, which may constitute part of or the whole set of adaptive mutations from an ALE. Negative correlations should represent features that result in neutral or negative effects on the strain's fitness when mutated together. The nature of selection for growth with ALE experiments results in more and stronger positive correlations than negative correlations between mutated features (Figure C.13,

Figure 4.4a). Hierarchical clustering of correlated features results in a distribution of cluster sizes with a median of 4 features (Figure 4.4a, Figure 4.4b), potentially describing the general need for only four mutated features to achieve a system optimization through ALE.

The ALE conditions metadata enables efforts in linking mutated features to conditions, isolating subsets of mutations from across experiments that may be related to conditions of interest. There are 93 unique conditions across 10 different condition types (Figure 4.2b, Figure C.1 through Figure C.9). Due to the variety of experiments consolidated in the data of this study, some conditions will contain more potential for designs than others. The supplement, carbon-source and strain conditions have substantially more associated features than the other condition types (Figure C.11), likely reflecting the variety of conditions and number of samples per condition type (Figure 4.2b, Figure C.1 through Figure C.9). Temperature has the highest median of associated unique features per condition (Figure C.12), with the other condition types nearly maintaining their previous ranking relative to each other. Operons and genes have the largest amount of associated conditions (Figure 4.4c). This is likely due to operons being composed of genes, which have the largest variety of uniquely mutated features in this dataset (Figure 4.3a). These associations can also describe the specificity of the mutated features for the given conditions. Most features across all types are associated with only a small set of conditions, though there exist some outliers that are associated with a broad range of conditions and could therefore be applicable to a broad range of stresses (Figure 4.4d). According to these trends, it is expected that variants derived from these associations are going to be gene-centric and only involve a small number of coding and non-coding features.

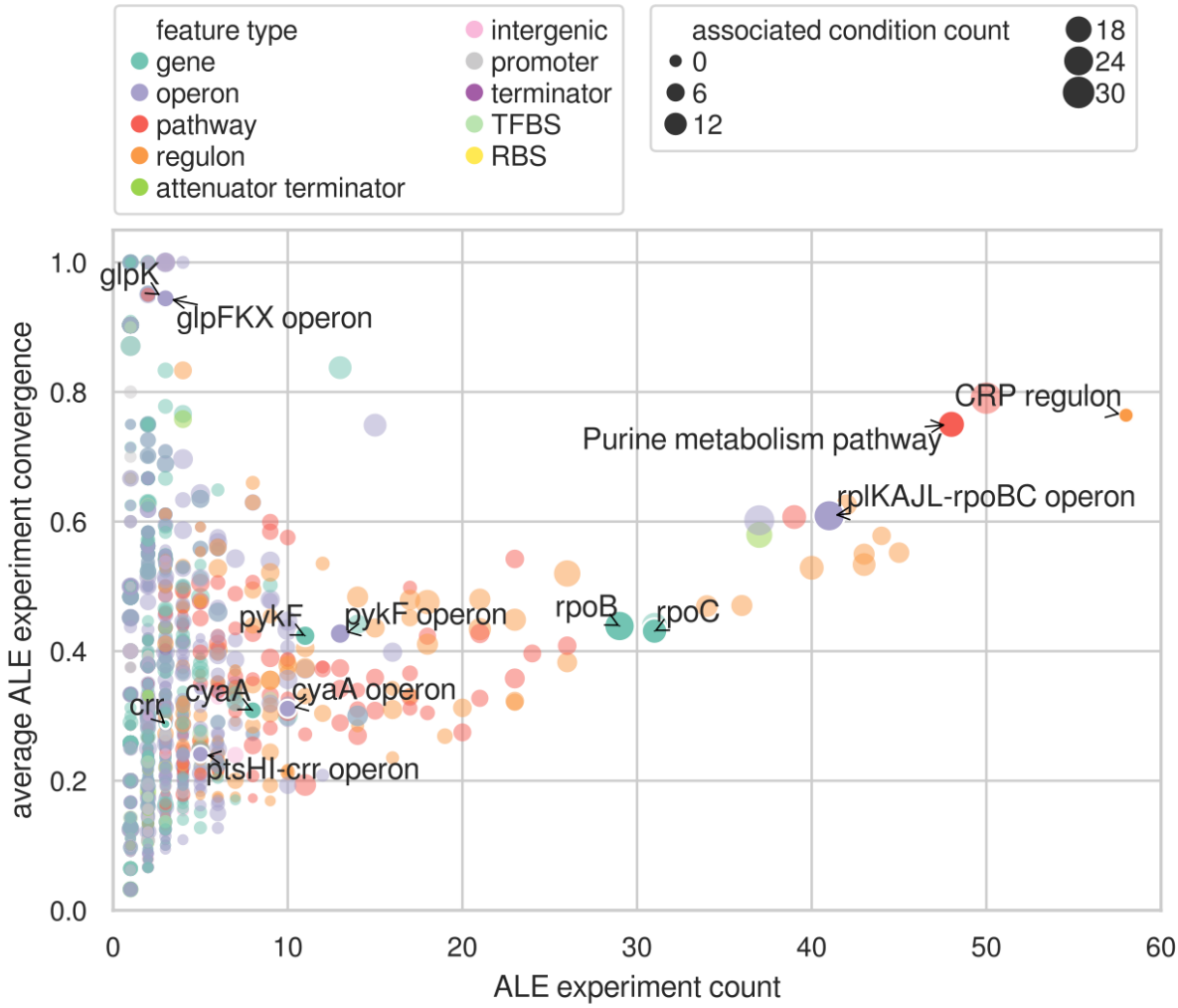


Figure 4.5: The combination of average convergence, ALE experiment count, and number of associated conditions for individual mutated features.

Some potentially beneficial mutations may only be apparent through the aggregation of ALE data

Some mutated features are more often selected across replicate ALEs of individual experiments and are described as having some measure of convergence [7] (Figure 4.5, Figure C.10). The phenomenon of convergence is leveraged in evolution experiments to identify key genomic features and sequence changes for the given conditions. The degree of convergence that mutated

features exhibit in their respective experiments provides a measure of selection strength for their mutation, suggesting an approximation for the degree of benefit from their mutation. Mutated features generally have low convergence (Figure C.10), though there are some that demonstrate high convergence and therefore strong evidence of selection (Figure 4.5, Figure C.10). Some features mutated in many experiments are observed with low convergence and associated conditions (Figure 4.5), indicating that their mutations may manifest in low amounts across experiments and are selected by very specific ALE conditions. These features may represent beneficial sequence changes that are only identifiable through the aggregation of ALE experiment mutations.

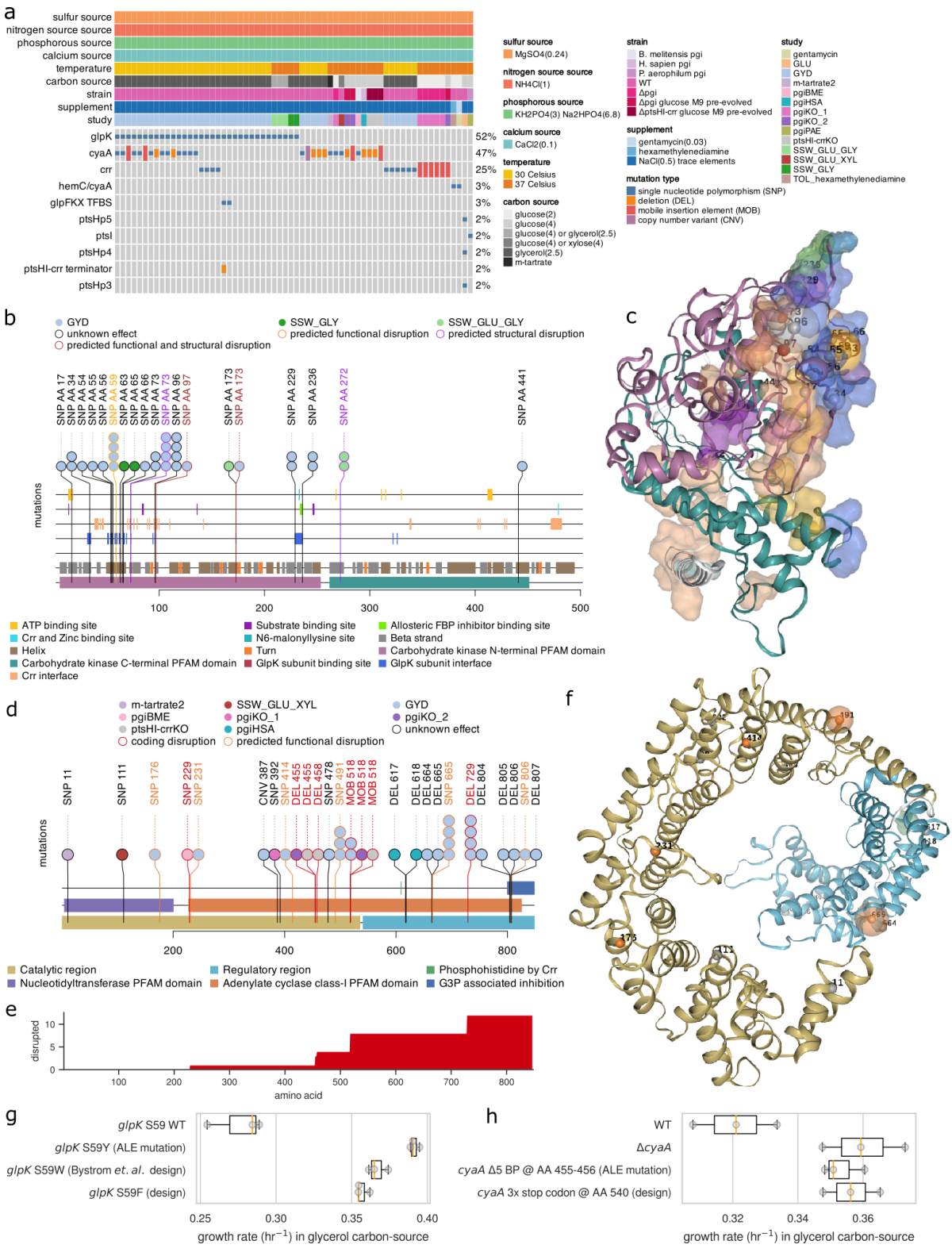
4.3.3 Case studies

The meta-analysis results reveal many opportunities for variant designs with the consolidated ALE dataset. The following sections describe the derivation of three variant designs using this ALE data and the workflow (Figure 4.1). Implementation of this workflow included additional annotations specific to the emergent genomic features of interest; these annotations are described in the Methods section.

An *E. coli* K-12 MG1655 strain design for glycerol as a carbon source

The use of alternative carbon sources for bioproduction could prove to be an important strategy in maintaining feedstock flexibility. Glycerol has been shown to serve as a feedstock for the production of valuable biochemicals, such as 1,3-PDO [13], ethanol [14], and limonene [15]. Glycerol-associated ALE mutations [7, 16] thus represent an opportunity for an ALE-derived strain design with possible applications.

Figure 4.6: (a) An oncoplot demonstrating the types of mutations to genomic features on operons of interest (operons *cyaA*, *glpFKX*, and *ptsHI-crr*) and the conditions for the ALE samples hosting these mutations. (b) A mutation needle plot for mutated amino acids across GlpK's amino acid chain. (c) GlpK's 3D structure and mutated residues from mutations that do not cause downstream coding disruptions. The residue chain and transparent surfaces are colored according to the legend of the corresponding mutation needle plot. Mutations are represented by a small opaque sphere with a value representing their amino acid position on the corresponding mutation needle plot. The color of the mutation's sphere corresponds to the mutation's predicted effect as described by the legend on the corresponding mutation need plot. The transparent sphere centered on the mutations' opaque sphere represents the number of mutations with a specific predicted effect on that position. The angle shown illustrates how all the GlpK GlpK interface surfaces are oriented on the same side of the 3D structure along with the clustering of mutations on or near these surfaces. (d) A mutation needle plot for mutated amino acids across CyaA's amino acid chain. (e) The accumulation of the disrupted amino acids downstream of coding disruptive (truncating) mutations from mutation needle plot. (f) CyaA's protein structure and mutated residues from mutations that do not cause downstream coding disruptions. (g) The growth rates of the mutants harboring ALE mutations and designed variants for GlpK in the selection pressure of glycerol as a carbon source. (h) The growth rates of the mutants harboring ALE mutations and designed variants for CyaA in the selection pressure of glycerol as a carbon source.


a
sulfur source
nitrogen source source
phosphorous source
calcium source
temperature
carbon source
strain
supplement
study
glpK
cyaA
crr
hemC/cyaA
glpFKX TFBS
ptsHp5
ptsI
ptsHp4
ptsHI-crr terminator
ptsHp3
52%
47%
25%
3%
3%
2%
2%
2%
2%
2%
sulfur source
MgSO4(0.24)
nitrogen source source
NH4Cl(1)
phosphorous source
KH2PO4(3) Na2HPO4(6.8)
calcium source
CaCl2(0.1)
temperature
30 Celsius
37 Celsius
carbon source
glucose(2)
glucose(4)
glucose(4) or glycerol(2.5)
glucose(4) or xylose(4)
glycerol(2.5)
m-tartrate
strain
B. melitensis pgi
H. sapien pgi
P. aerophilum pgi
WT
Δpgi
Δpgi glucose M9 pre-evolved
ΔptsHI-crr glucose M9 pre-evolved
supplement
gentamycin(0.03)
hexamethylenediamine
NaCl(0.5) trace elements
mutation type
single nucleotide polymorphism (SNP)
deletion (DEL)
mobile insertion element (MOB)
copy number variant (CNV)
study
gentamycin
GLU
GYD
m-tartrate2
pgiBME
pgiHSA
pgiKO_1
pgiKO_2
pgiPAE
ptsHI-crrKO
SSW_GLU_GLY
SSW_GLU_XYL
SSW_GLY
TOL_hexamethylenediamine
b
GYD
unknown effect
predicted functional and structural disruption
SSW_GLY
predicted functional disruption
SSW_GLU_GLY
predicted structural disruption
SNP AA 17
SNP AA 34
SNP AA 54
SNP AA 55
SNP AA 56
SNP AA 59
SNP AA 63
SNP AA 65
SNP AA 66
SNP AA 73
SNP AA 73
SNP AA 96
SNP AA 97
SNP AA 173
SNP AA 173
SNP AA 229
SNP AA 236
SNP AA 272
SNP AA 441
mutations
100
200
300
400
500
ATP binding site
Crr and Zinc binding site
Helix
Carbohydrate kinase C-terminal PFAM domain
Crr interface
Substrate binding site
N6-malonyllysine site
Turn
GlpK subunit binding site
Allosteric FBP inhibitor binding site
Beta strand
Carbohydrate kinase N-terminal PFAM domain
GlpK subunit interface
c
d
m-tartrate2
pgiBME
ptsHI-crrKO
coding disruption
SSW_GLU_XYL
pgiKO_1
pgiHSA
predicted functional disruption
GYD
pgiKO_2
unknown effect
SNP 11
SNP 111
SNP 176
SNP 229
SNP 231
CNV 387
SNP 392
SNP 414
DEL 455
DEL 455
DEL 458
SNP 478
SNP 491
MOB 518
MOB 518
MOB 518
DEL 617
DEL 618
DEL 664
DEL 665
SNP 665
DEL 729
DEL 804
DEL 805
DEL 806
SNP 806
DEL 807
mutations
200
400
600
800
Catalytic region
Nucleotidyltransferase PFAM domain
Regulatory region
Adenylate cyclase class-I PFAM domain
Phosphohistidine by Crr
G3P associated inhibition
f
e
disrupted
10
5
0
100
200
300
400
500
600
700
800
amino acid
g
glpK S59 WT
glpK S59Y (ALE mutation)
glpK S59W (Bystrom et. al. design)
glpK S59F (design)
0.25
0.30
0.35
0.40
growth rate (hr−1) in glycerol carbon-source
h
WT
ΔcyaA
cyaA Δ5 BP @ AA 455-456 (ALE mutation)
cyaA 3x stop codon @ AA 540 (design)
0.32
0.34
0.36
growth rate (hr−1) in glycerol carbon-source

Associations with conditions and analysis of multi-scale mutation annotations outlines systems important for adaptation to glycerol as a carbon source

Within the dataset, 144 mutations had at least one of their host features significantly associated with glycerol as a carbon source (Figure C.14). The CRP regulon hosts the most mutated features for ALEs using glycerol as a carbon source (Table C.7), is associated with this selection pressure, and is one of the most frequently mutated regulons across this dataset's ALE experiments (Figure 4.5, Figure C.16). The CRP regulon describes many operons that encode for catabolic functions, including secondary carbon source metabolism. The CRP regulon is statistically associated with six different conditions, three of which are secondary carbon sources (Figure C.16). The three most frequently mutated operons associated with glycerol as a carbon source and linked to CRP are *glpFKX*, *cyaA*, and *ptsHI-crr* (Table C.6, Figure C.15, Figure C.16). The genes *glpK*, *cyaA*, and *crr* are the most frequently mutated features of their operons (Figure C.15, Figure 4.6a). Mutations to the CRP regulon, *glpFKX* operon, and *glpK* are strongly selected for (Figure 4.5) and *glpFKX* mutations are strongly associated with conditions involving glycerol and a temperature of 30 Celcius (Table C.5, Figure C.16). While the *cyaA* and *ptsHI-crr* operons and the *cyaA* and *crr* genes are less strongly selected for by their ALE experiments, they are still seen mutated in multiple ALE experiments (Figure 4.5) and are associated with a partially overlapping set of conditions (Figure C.16). Multiple strong associations to ALE conditions suggest the potential that mutations to these features could be beneficial for a broad set of stresses. Mutations to *glpK* and *cyaA* or *glpK* and *crr* are found in the same samples, though mutations to *crr* and *cyaA* are not (Figure 4.6a). This, along with correlations between the mutated genomic features associated with glycerol as a carbon source (Figure C.29), suggest that mutations to *crr* and *cyaA* have a negative epistatic relationship [7].

Mutations to GlpK strongly cluster around and on subunit interface and binding sites

Among *glpK*, *cyaA*, and *crr*, *glpK* was the most frequently mutated. Further, it was observed to mutate in ALE experiments that involve substrate switching between glucose and glycerol (Figure 4.6a); all ALE experiment mutations to *glpK* were investigated to understand if mutation types clustered within the gene according to conditions. All mutations to *glpK* were SNPs and frequently landed in codons for amino acids involved in GlpK's subunit interface (Figure 4.6b), where GlpK can form both a dimer and tetramer. The subunit interface amino acids are spread across GlpK's sequence (Figure 4.6b), though in the 3D structure model (3EZW) their residue surfaces all group together to face the same direction (Figure 4.6c), revealing further clustering of mutations specific to 3D space. Finding the distances between mutated residues and GlpK features according to their 3D positions on GlpK's structure (3EZW) provided for a potentially more accurate measure of nearness between mutations and features. GlpK subunit interfaces continue to be nearest to or directly host the most SNPs (Figure C.17), though the mutations to the GlpK subunit binding sites have the highest proportion of mutations with a predicted effect (Figure C.18). 4/5 of the SNPs to the GlpK subunit binding sites are the functionally disruptive (SIFT score < 0.05) SNP S59Y. GlpK (glycerol kinase) is part of the pathway that utilizes glycerol as a carbon source. GlpK can form a catalytically active homotetramer and homodimer, though the homotetramer can be allosterically inhibited by fructose-1,6-bisphosphate, a downstream product serving as negative feedback. Mutations to Serine 59 have already been shown to disable homotetramer formation through steric incompatibility [17]. This leaves the homodimer formation, which may accomplish a higher overall rate of glycerol metabolism due to the lack of inhibition.

The GlpK variant design targets the subunit interface binding site

Across all possible amino acid substitutions, SNP S59Y is highly ranked according to SIFT scores, size, and flexibility difference relative to wild-type residues (SFigure C.19). The residue substitution of tryptophan (W) scores higher or similar to the tyrosine (Y) ALE substitution in the previously mentioned categories, as well as being categorized as having a bulky aromatic sidechain (Figure C.19b), thus making it a good candidate for a derived design. A substitution of tryptophan at this position has already been characterized to eliminate inhibition of GlpK's catalytic activity [17]. Since the tryptophan substitution has already been characterized, another novel substitution was pursued for the purposes of this study. Phenylalanine (F) is the only other residue that is both characterized as being bulky and having an aromatic sidechain as well as scoring similarly to tyrosine and tryptophan substitutions. The final proposed design for GlpK was that of S59F.

GlpK mutant and designs outgrow wildtype with glycerol as a carbon source

To examine the fitness changes from mutants relative to wildtype with glycerol as a carbon source, growth screens were performed on manually reconstructed strains harboring the S59Y ALE mutation along with the designed S59W or S59F GlpK variants. The results show that the mutant and designs have similar growth rates as well as higher growth rates than wildtype (Figure 4.6g).

A partial truncation of CyaA may be beneficial for a broad set of conditions

cyaA was the second most frequently mutated gene of those associated with glycerol and its mutations were often found in samples with mutations to *glpK* (Figure 4.6a). *cyaA* was mu-

tated in multiple experiments of different conditions (Figure 4.6a); all ALE experiment mutations to *cyaA* were investigated to understand if mutation types clustered within the gene according to conditions. Mutations from multiple experiments seem to cluster near the middle of the amino acid sequence (Figure 4.6d). Many of the mutations within this cluster cause a frameshift or truncation, thereby disrupting the downstream coding sequence. The accumulation of downstream disrupted amino acids demonstrated that the regulatory region was often targeted for disruption across varying ALE conditions, including that of glycerol as a carbon source (Figure 4.6e). Observing the mutated residues on *cyaA*'s protein structure from mutations other than coding disruptions, there was no novel 3D clustering of mutated residues not already apparent with the linear analysis (Figure 4.6f). This cluster of mutations from ALE experiments with various selection pressures suggests that a specific change to *cyaA* may have broad applicability.

A partial truncation of CyaA may be maintaining carbon catabolite repression

CyaA (adenylate cyclase) is part of the pathway that generates the activated CRP complex (cAMP-CRP), which goes on to activate genes for multiple secondary carbon source catabolic systems [18]. CyaA's regulatory region is thought to activate the enzyme [19]. For the condition of glycerol as a carbon source, CyaA would become activated and produce cAMP, therefore activating these catabolic systems [19]. A truncation of CyaA has been shown to prevent cAMP production [20] and the downstream activation of CRP is expected to be nullified. The repression of these secondary carbon source catabolic systems is known as Carbon Catabolite Repression (CCR) and is normally enforced by the phosphotransferase system in the presence of glucose [18]. With glycerol as a carbon source, ALE evolved strains may have found a way to maintain CCR through mutations to *cyaA* while still allowing for the activation of glycerol metabolism.

CCR maintenance may ultimately enable a more efficient cell metabolism, since it represses the activation of multiple unnecessary catabolic systems.

Crr mutations may also be working to maintain carbon catabolite repression

crr was the third most frequently mutated gene of those associated with glycerol (Figure 4.6a) and its mutations were often found in samples with mutations to *glpK* and never found in samples with mutations to *cyaA* (Figure 4.6a). The absence of mutated *cyaA* and *crr* genes in the same sample has been hypothesized to be due to their mutations having similar effects to the same system in the case of ALEs with a glycerol carbon source [7]. *crr* was mutated in multiple experiments of different conditions (Figure 4.6a); all ALE experiment mutations to *crr* were investigated to understand if mutation types clustered within the gene according to selection pressure. Most mutations to *crr*'s amino-acid sequence fall on or near interfaces. These interfaces describe separate surfaces of Crr's structure [21] (Figure C.21). Finding the distances between mutated residues and Crr features according to their 3D positions on Crr's structure [21] provided for a potentially more accurate measure of nearness between mutations and features. Crr feature residues truncated by an upstream coding disruption are additionally included. The Crr Crr interface hosts the most mutated residues (Figure C.22), though its mutation type differs from those mutated residues to all other features (Figure C.23). Mutations to the Crr Crr interface are also specific to Δpgi ALEs while mutations to all other features are specific to glycerol carbon source ALEs (Figure C.20). Mutations from glycerol carbon source ALEs additionally cluster near each other on or near the same surface of the 3D structure (Figure C.21). This surface hosts the GlpK, PtsG, PtsH, PtsI, and FrsA interfaces as well as binding and active sites. Unphosphorylated Crr binds to and inhibits GlpK in vitro [18], though in the case of a glycerol

as a carbon source, phosphorylated Crr should be more abundant. Crr interacts with CyaA and will activate CyaA's cAMP production if phosphorylated [18]. Though this work does not include CyaA interface data for Crr, the phosphoryl group active sites are located in the middle of the multi-interface residues shared with GlpK; mutations to this interface surface may prevent the interaction with CyaA that activates cAMP production. All but one mutation clustering near this interface were predicted to be disruptive (Figure C.21). The mutations to Crr that manifest in the condition of glycerol as a carbon source may be working to maintain CCR, similar to those mutations to CyaA.

A CyaA variant is chosen to maintain carbon catabolite repression in the presence of glycerol

Mutations to *cyaA* have a more interpretable effect on the gene than those to Crr: the truncation of CyaA's regulatory region. The 5 bp deletion starting within amino acid 455 in CyaA is chosen as the representative ALE mutation for CCR maintenance due to it being the most easily reintroducible truncation to the entire regulatory region. To deactivate CyaA's in a similar manner to the ALE mutations, the proposed CCR maintenance strain design was the insertion of 3 stop codons at amino acid 540, immediately upstream of the regulatory region.

CyaA mutant and designs outgrow wildtype with glycerol as a carbon source

To examine the fitness changes from mutants relative to wildtype, growth screens were performed on the reconstructed strains harboring the *cyaA* ALE mutation along with the designed truncation. The results show the mutant and designs have similar growth rates as well as higher growth rates than wildtype (Figure 4.6h).

An *E. coli* K-12 MG1655 strain design for high concentrations of isobutyric acid

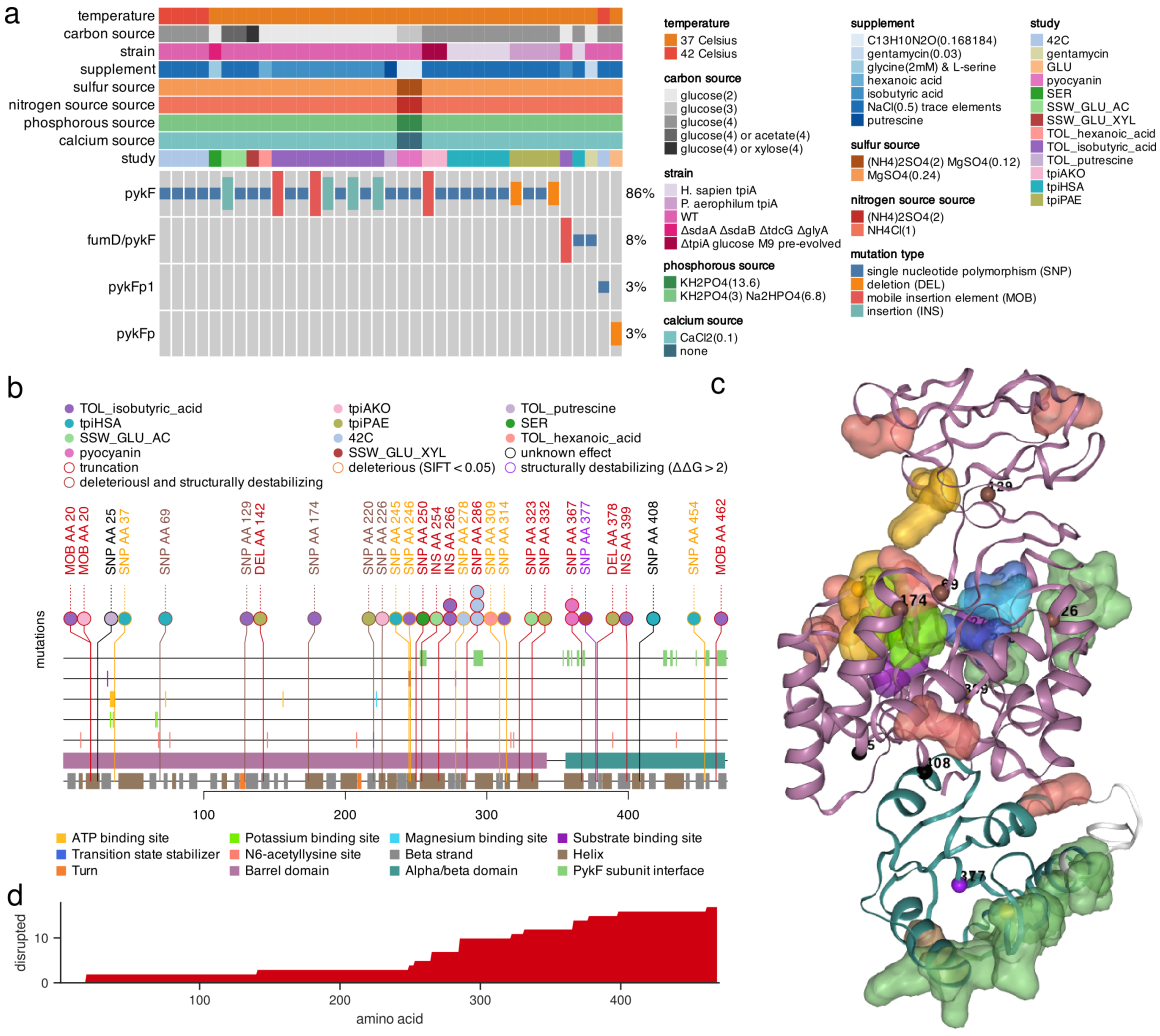
Tolerance is a key phenotype for microbial cell factories. Industrial requirements may have strains exposed to or produce toxic concentrations of substrates or products. Genotypes that provide any improved tolerance to detrimental conditions can be valuable in that their tolerance can translate to higher concentrations of product, especially with large scale operations [22]. Isobutyrate is a biochemical with a market size of 100,000 tons in 2011 [23] and can be produced with an engineered *E. coli* strain [22]. Isobutyrate associated ALE mutations [22] represent an opportunity for an ALE-derived strain design with possible applications.

Associations with conditions and analysis of multi-scale mutation annotations outlines systems important for adaptation to isobutyric acid tolerance

Within the dataset, 73 mutations had at least one of their host features significantly associated with the condition of toxic isobutyrate concentrations (Figure C.24). The purine metabolism pathway hosts the most mutations ((Table C.10)), with the majority of the mutations coming from the *pykF* and *rplKAJL-rpoBC* operons (Figure C.25). Of these two, the *pykF* operon is strongly associated with toxic isobutyrate concentrations (Figure C.28). The purine metabolic pathway, *rplKAJL-rpoBC* operon, *pykF* operon, and *pykF* are frequently mutated features across this dataset's ALE experiments (Figure 4.5). The purine metabolic pathway and *rplKAJL-rpoBC* operon are strongly selected for in ALE experiments, while mutations to *pykF* and its operon are less so on average Figure 4.5). *pykF* and its operon are also more specific in their associations to conditions, where the *rplKAJL-rpoBC* operon and purine metabolic pathway mutation associations are more broad (Figure 4.5, Figure C.28). Correlations between frequently mutated genomic features demonstrated *pykF* and *rpoB* as positively correlated (Table C.8, Fig-

ure C.30). The positive correlation between *pykF* and *rpoB* along with their differing associations indicate that these compatible mutations are adapting for different selection pressures.

Figure 4.7: (a) An oncoplot demonstrating mutations linked to the pykF operon across all ALE experiments of this study's data. (b) A mutation needle plot for mutated amino acids across PykF's amino acid chain. The plot below the mutation needle plot represents the accumulation of the disrupted amino acids downstream of truncation mutations from the above plot. (c) PykF's 3D structure and mutated residues. No truncating mutations are included. The residue chain and transparent surfaces are colored according to the legend of the corresponding mutation needle plot. Mutations are represented by a small opaque sphere with a value representing their amino acid position on the corresponding mutation needle plot. The color of the mutation's sphere corresponds to the mutation's predicted effect as described by the legend on the corresponding mutation need plot. The transparent sphere centered on the mutations opaque sphere represents the number of mutations with a specific predicted effect on that position. The angle shown illustrates how most of the mutations cluster in 3D space around the area which hosts most of the catalytic domains. (d) The accumulation of the disrupted amino acids downstream of coding disruptive (truncating) mutations from mutation needle plot. (e) The growth rates of WT, a $\Delta pykF$ strain, the *pykF* ALE mutant, and the *pykF* designed variant with toxic concentration of isobutyrate (12.5 g/L). A $\Delta pykF$ strain mutant was used to investigate for any difference between the strains that partially truncate *pykF* and its full truncation. (f) The growth rates of WT, a $\Delta pykF$ strain, the *pykF* ALE mutant, and the *pykF* designed variant with glucose as a carbon source. A $\Delta pykF$ strain mutant was used to investigate for any difference between the strains that partially truncate *pykF* and its full truncation.


a
temperature
carbon source
strain
supplement
sulfur source
nitrogen source source
phosphorous source
calcium source
study
pykF
fumD/pykF
pykFp1
pykFp
86%
8%
3%
3%
temperature
37 Celsius
42 Celsius
carbon source
glucose(2)
glucose(3)
glucose(4)
glucose(4) or acetate(4)
glucose(4) or xylose(4)
strain
H. sapien tpiA
P. aerophilum tpiA
WT
ΔsdaA ΔsdaB ΔtdcG ΔglyA
ΔtpiA glucose M9 pre-evolved
phosphorous source
KH2PO4(13.6)
KH2PO4(3) Na2HPO4(6.8)
calcium source
CaCl2(0.1)
none
supplement
C13H10N2O(0.168184)
gentamycin(0.03)
glycine(2mM) & L-serine
hexanoic acid
isobutyric acid
NaCl(0.5) trace elements
putrescine
sulfur source
(NH4)2SO4(2) MgSO4(0.12)
MgSO4(0.24)
nitrogen source source
(NH4)2SO4(2)
NH4Cl(1)
mutation type
single nucleotide polymorphism (SNP)
deletion (DEL)
mobile insertion element (MOB)
insertion (INS)
study
42C
gentamycin
GLU
pyocyanin
SER
SSW_GLU_AC
SSW_GLU_XYL
TOL_hexanoic_acid
TOL_isobutyric_acid
TOL_putrescine
tpiAKO
tpiHSA
tpiPAE
b
TOL_isobutyric_acid
tpiHSA
SSW_GLU_AC
pyocyanin
truncation
deleteriousl and structurally destabilizing
tpiAKO
tpiPAE
42C
SSW_GLU_XYL
deleterious (SIFT < 0.05)
TOL_putrescine
SER
TOL_hexanoic_acid
unknown effect
structurally destabilizing (ΔΔG > 2)
MOB AA 20
MOB AA 20
SNP AA 25
SNP AA 37
SNP AA 69
SNP AA 129
DEL AA 142
SNP AA 174
SNP AA 220
SNP AA 226
SNP AA 245
SNP AA 246
SNP AA 250
INS AA 254
INS AA 266
SNP AA 278
SNP AA 286
SNP AA 309
SNP AA 314
SNP AA 323
SNP AA 332
SNP AA 367
SNP AA 377
DEL AA 378
INS AA 399
SNP AA 408
SNP AA 454
MOB AA 462
mutations
100
200
300
400
ATP binding site
Transition state stabilizer
Turn
Potassium binding site
N6-acetyllysine site
Barrel domain
Magnesium binding site
Beta strand
Alpha/beta domain
Substrate binding site
Helix
PykF subunit interface
c
d
disrupted
10
0
100
200
300
400
amino acid
e
WT
ΔpykF
pykF 1 BP INS @ AA 266 (ALE mutation)
pykF 3x stop codon @ AA 253 (design)
0.0
0.1
0.2
0.3
growth rate (hr−1) with isobutyrate(12.5 g/L)
f
WT
ΔpykF
pykF 1 BP INS @ AA 266 (ALE mutation)
pykF 3x stop codon @ AA 253 (design)
0.68
0.70
0.72
0.74
0.76
growth rate (hr−1) in glucose carbon-source

Mutations to PykF have two trends depending on how they affect the gene product

The *pykF* operon is frequently mutated across this study's many ALE experiments. These mutations were gathered and investigated to understand if mutations clustered according to ALE experiment conditions. Most of these mutations target *pykF*'s coding sequence, with some to the operon's non-coding features (Figure **??**a). A mutation to the *pykF* operon's non-coding features seems to preclude a mutation to any of the others, suggesting that they have related effects on the host strain (Figure **??**a). Mutations to *pykF* were spread across its sequence, with the majority contained in the second half of the sequence (Figure **??**b). Due to the large variety of experiments mutating *pykF*, the broad distribution of mutations across *pykF*'s sequence, and the large amount and variety of structural feature annotations available for PykF, it is expected that the trends described by the aggregate of all ALE experiment mutations would be most revealing. Many of the mutations to *pykF* were predicted to disrupt the coding sequence through truncations or frameshifts, or disrupt the potential function and/or structural stability of a domain through a non-synonymous amino acid substitution. Considering the accumulation of disrupted amino acids downstream of a truncation or frameshift, the PykF tetramer subunit interfaces were the most frequently mutated features on PykF (Figure **??**b, Figure C.26). These coding disruption mutations also seem to cluster on the second half of PykF's sequence, resulting in the downstream coding disruptions primarily affecting the PykF tetramer subunit interfaces while avoiding active sites in the first half of the sequence (Figure **??**b). Those mutations predicted not to cause truncations or frameshifts were found on PykF's 3D structure nearest to binding sites within the cleft of PykF's barrel domain (Figure C.17, Figure **??**c). All mutations to PykF and those specifically from the toxic isobutyrate ALEs follow the same trend across features on PykF's structure (Figure C.26), providing evidence that mutations to *pykF* across different

sets of conditions may accomplish similar outcomes. These trends suggest unique functional targets for the different mutation types, with those of the coding disruptive mutations having the clearest outcome: a disruption to PykF's ability to form a complex.

PykF truncations may be beneficial to glucose uptake

PykF mutations are hypothesized to allow for increased concentrations of phosphoenolpyruvate (PEP) by removing one of the reactions that converts PEP to pyruvate, that of pyruvate kinase (PYK). In glucose minimal media, more PEP could result in a higher rate of glucose uptake through the phosphotransferase system [24]. Full truncations of *pykF* are thought to be valuable with scarce glucose [25], where partial truncations may be more valuable in abundant glucose [24] due to the possible maintenance of some pyruvate production capability. All of this study's *pykF* mutations manifested in strains using glucose as a carbon source (Figure **??**a) and the majority of the truncations occurred near the middle of the coding sequence, where truncations avoided catalytic sites encoded in the first half of the gene (Figure **??**b). The chosen representative ALE mutation may therefore benefit the host strain when using glucose as a carbon source.

The PykF variant design objective is clarified through aggregated ALE data

A truncating insertion to amino acid 266 was chosen as the representative ALE mutation to PykF for the toxic isobutyrate condition since it represents a clear trend in disrupting the PykF subunit interfaces and manifested in two different ALEs with this selection pressure. The interpretable trend of disrupted amino acids accumulating across PykF subunit interfaces led to a variant design that truncated all of the PykF subunit interfaces: an insertion of 3 stop

codons at amino acid 253, immediately upstream of all the PykF subunit interfaces. These ALE mutations and the derived genome design were initially inspired by their strong selection with the isobutyrate tolerance ALEs, though their benefit may instead be directly derived from growth on an abundance of glucose as a carbon source. All ALEs with mutations to *pykF* include abundant glucose as a carbon source (FFigure **??**a); these mutations may have contributed to the fitness of their host, though this fitness may have been secondary to other mutations with higher fitness for the primary selection pressure. These more subtle optimizations become more clear through the meta-analysis of the consolidated ALE dataset.

PykF mutant and designs outperform wildtype with toxic isobutyrate levels though don't have substantial effect with only glucose

Growth screens were performed on the reconstructed strains harboring the *pykF* ALE mutation along with the designed truncation to examine the differences in fitness between mutants and wildtype. Growth screens for both toxic isobutyrate levels and competitive glucose uptake were performed. The results show the mutant and designs have similar growth rates as well as higher growth rates than wildtype with toxic isobutyrate concentrations (Figure **??**e) while not demonstrating a substantial difference between mutants and wildtype with only glucose (Figure **??**f).

4.4 Discussion

The phenomena of convergent mutations with experimental evolution serves well to reveal genotypic changes that likely result in a higher level of fitness for the host strain. Experimental evolution methods such as ALE leverage convergence to hypothesize which mutations bring about

fitness benefits. ALE has been successfully applied in strain engineering, though their resulting mutation sets can be too small to deduce the functions in which the ALE mutations are targeting. This work demonstrates that enriching mutation data, experimental design metadata, and functional annotations by leveraging a variety of public resources amplified ALE mutation trends, resulting in cases where the target biological functions mechanisms upon by ALE mutations could be deduced.

Multiple methods were implemented to find ALE mutation trends at different levels of detail. These methods were organized into a workflow so that the sequential execution of steps leads to the design of sequence variants. Associations between mutated features and conditions extracted the mutated features potentially relevant for a condition of interest in strain design. Multi-scale annotations helped group mutated features that may have similar effects. The meta-analysis of multiple dimensions suggests that most ALE-derived genotypic solutions to a stress will require a small number of sequence changes, and those sequence changes will most likely be in genes. This may be due to many factors, one in which is that the genome of E. coli K-12 MG1655 is mostly composed of coding sequences [26].

The case studies in this work demonstrate that enriched ALE data and methods provide enough evidence to deduce the sequence change properties and resulting molecular mechanisms targeted by ALE mutations. From these results, sequence-based strain designs were derived and revealed further insights into ALE-data-driven strain design. Mutation trends in genes became more apparent through the combination of mutation effect predictions and increased spatial dimensionality. Trends involving truncating mutations were interpretable when mutations and the affected downstream amino acids were mapped onto the annotated 1D amino acid sequence. Trends involving non-truncating mutations were more clear and interpretable when mutations

were mapped onto 3D structures, since 3D structures better represent the spatial dimensions of a gene product. Of the two mutation trends to pykF, one was revealed to involve structurally disruptive mutations clustering around a 3D structure forming a catalytic pocket. Analytics of mutation trends characterized different types of adaptive mutation strategies on genes, some of which were clear enough to derive novel variants from. Two different types of variants were derived: 1) the loss of molecular function through truncations and 2) a change in binding ability through amino acid substitution. The variant designs of this work ultimately targeted the reduction of functionality encoded within a gene. These designs and their original mutations may be trading the robustness of a system for more simple and higher performing processes.

As shown with the computational results from the glycerol carbon-source case study, designs could also result from a combined set of compatible variants, where each variant optimizes for a separate stress. A strain's environment is naturally composed of multiple conditions, therefore compatible optimizations would be valuable in addressing multi-stress circumstances such as those of industrial scale fermentation (physical, chemical, and biological) [27–29].

The ALE and designed mutant strains of this work were screened for their phenotypes. The mutant strains were found to be more fit than wildtype in the conditions that initially selected for their presence in ALEs. The ability of the strains hosting designed variants to achieve similar growth rates to beneficial ALE mutants demonstrates the possibility to design variants derived from aggregated ALE data. The similarity between ALE and designed mutant growth rates also suggests that ALE processes of the current scale and time-span represented in this work may not find all possible beneficial sequence changes. This lack of full coverage for beneficial sequence changes by ALE processes may be due to the probability of specific mutational sequence changes. For example, GlpK's ALE mutation involved a single base pair substitution, while both

its designed variants involved two base pair substitutions within the same codon. The sequence changes involved in the designed variants may be much less likely to occur with ALE than the ALE mutation, which manifested in 4 of 35 replicate ALEs using glycerol as a carbon source. Thus, there may exist potential for beneficial variants not revealed through ALE.

In the cases involving truncations, ALE mutants and designed strains didn't produce more benefit than the corresponding gene knockouts. This was expected for CyaA, as the truncations targeted a regulatory region thought to activate CyaA [19]. Previous studies proposed that partial PykF truncations may modulate its activity [24], which was expected to be more beneficial than a full truncation in the case of abundant glucose. The growth screens demonstrated no substantial difference in fitness between wildtype, and both full and partial truncations of PykF with abundant glucose. The modulation of protein activity through partial truncation remains hypothetically possible, where the partial truncation could remove an inhibitory domain, though the growth screen results demonstrated this as not the case with PykF.

The methods in this work demonstrate the value in leveraging diverse public resources to describe ALE variants, but are not without limitations. The success of adequately characterizing the beneficial mechanisms of mutations relies upon the availability of genome annotations local to the mutations of interest as well as tools that can describe the nature and magnitude of a mutation's severity. The efforts of this work leveraged an already available whole-genome multi-scale annotation framework and further enriched mutation annotations local to features genes of interest. These additional annotations, coming from resources such as Uniprot [30], EcoCyc [31], Pfam [32], and Mutfunc [33], are available for the whole genome, though efforts have not yet been made to consolidate them into a unified computational resource. Additionally, specific mutation characterization greatly benefitted from investigating the other mutations local to a

feature of interest. Variants can therefore serve as an additional genome annotation type, and are especially useful if described by the conditions they were found in. There also exists the potential to grow this set of consolidated mutations with those from other studies and with natural variants. Traditional genome annotations typically don't include variants, though recent efforts have developed a bioinformatic resource that combines genome annotations as well as variants: the Bitome [34]. The Bitome could serve as the locus for consolidating whole-genome multi-scale feature annotations along with variants and their metadata for a comprehensive, high-resolution, genomic resource per organism.

4.5 Conclusion

Three novel variants applicable to strain designs have resulted from this work: two variant designs for glycerol metabolism and a variant design for isobutyrate tolerance. These designs demonstrate how the meta-analysis of aggregated public ALE sequence-level mutation data, rich annotations, and structural biology methods can provide enough evidence to interpret ALE mutation objectives towards strain design. Besides serving as proofs-of-concept, the case study designs may hold value for applications in microbial cell factories and beyond. Until rational methods can predict all possible biological paths between genotypes and phenotypes, data-driven methods will continue to provide value towards bioengineering when combined with consolidated public data such as demonstrated here.

Acknowledgements

Conceptualization: PVP, DZ, BOP, AMF. Technical support: MW, SB, CD, JJ, LY, SHK. Conceptual support: JTY, DZ, AMF, BOP, CD. Writing: PVP, JTY, DZ, SB, CD.

The authors gratefully acknowledge Robin Cai, Douglas McCloskey, and Emre Ozdemir for their help with the technologies involved in this project and Marc Abrams for his editorial comments. The authors gratefully acknowledge Connor Olson for his support with the characterization experiment work.

This work was funded by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (grant number NNF10CC1016517). Funding was also provided by the NIH National Institute of Allergy and Infectious Diseases (grant number U01AI124316).

Chapter 4 is a reprint of material published in: **Patrick V. Phaneuf**, Daniel C. Zielinski, James T. Yurkovich, Josefin Johnsen, Richard Szubin, Lei Yang, Se HyeuK Kim, Christopher Dalldorf, Muyao Wu, Sebastian Schulz, Bernhard O. Palsson, Adam M. Feist "Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data." The dissertation author is the primary author.

4.6 References

1. Garcia-Granados, R., Lerma-Escalera, J. A. & Morones-Ramirez, J. R. Metabolic Engineering and Synthetic Biology: Synergies, Future, and Challenges. en. *Front Bioeng Biotechnol* **7,** 36. ISSN: 2296-4185 (Mar. 2019).

2. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. en. *Cell* **164,** 1185–1197. ISSN: 0092-8674, 1097-4172 (Mar. 2016).

3. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. en. *Nat. Biotechnol.* **33,** 1061–1072. ISSN: 1087-0156, 1546-1696 (Oct. 2015).

4. Sandberg, T. E., Salazar, M. J., Weng, L. L., Palsson, B. O. & Feist, A. M. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. en. *Metab. Eng.* ISSN: 1096-7176, 1096-7184 (Aug. 2019).

5. Wang, X., Zorraquino, V., Kim, M., Tsoukalas, A. & Tagkopoulos, I. Predicting the evolution of Escherichia coli by a data-driven approach. en. *Nat. Commun.* **9,** 3562. ISSN: 2041-1723 (Sept. 2018).

6. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* **47,** D1164–D1171. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

7. Phaneuf, P. V., Yurkovich, J. T., Heckmann, D., Wu, M., Sandberg, T. E., King, Z. A., Tan, J., Palsson, B. O. & Feist, A. M. Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity. en. *BMC Genomics* **21,** 514. ISSN: 1471-2164 (July 2020).

8. Rudrapatna, V. A. & Butte, A. J. Open data informatics and data repurposing for IBD. en. *Nat. Rev. Gastroenterol. Hepatol.* **15,** 715–716. ISSN: 1759-5045, 1759-5053 (Dec. 2018).

9. Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortes-Ciriano, I., Zhou, D. C., Liang, W.-W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Van Allen, E. M., Liang, H., MC3 Working Group, Cancer Genome Atlas Research Network, Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R. & Ding, L. Comprehensive Characterization of Cancer Driver Genes and Mutations. en. *Cell* **173,** 371–385.e18. ISSN: 0092-8674, 1097-4172 (Apr. 2018).

10. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nat. Commun.* **9,** 4306. ISSN: 2041-1723 (Oct. 2018).

11. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. en. *Nat. Commun.* **5,** 4846. ISSN: 2041-1723 (Sept. 2014).

12. Zielinski, D. C., Patel, A. & Palsson, B. O. The Expanding Computational Toolbox for Engineering Microbial Phenotypes at the Genome Scale. en. *Microorganisms* **8,** 2050 (Dec. 2020).

13. Clomburg, J. M. & Gonzalez, R. Biofuel production in Escherichia coli: the role of metabolic engineering and synthetic biology. en. *Appl. Microbiol. Biotechnol.* **86,** 419–434. ISSN: 0175-7598, 1432-0614 (Mar. 2010).

14. Shams Yazdani, S. & Gonzalez, R. Engineering Escherichia coli for the efficient conversion of glycerol to ethanol and co-products. en. *Metab. Eng.* **10,** 340–351. ISSN: 1096-7176, 1096-7184 (Nov. 2008).

15. Willrodt, C., David, C., Cornelissen, S., Bühler, B., Julsing, M. K. & Schmid, A. Engineering the productivity of recombinant Escherichia coli for limonene formation from glycerol in minimal media. en. *Biotechnol. J.* **9,** 1000–1012. ISSN: 1860-6768, 1860-7314 (Aug. 2014).

16. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Appl. Environ. Microbiol.* **83** (July 2017).

17. Bystrom, C. E., Pettigrew, D. W., Branchaud, B. P., O'Brien, P. & Remington, S. J. Crystal structures of Escherichia coli glycerol kinase variant S58–¿W in complex with nonhydrolyzable ATP analogues reveal a putative active conformation of the enzyme as a result of domain motion. en. *Biochemistry* **38,** 3508–3518. ISSN: 0006-2960 (Mar. 1999).

18. Gorke, B. & Stulke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. en. *Nat. Rev. Microbiol.* **6,** 613–624. ISSN: 1740-1526, 1740-1534 (Aug. 2008).

19. Saier Jr, M. H., Kukita, C. & Zhang, Z. Transposon-mediated directed mutation in bacteria and eukaryotes. en. *Front. Biosci.* **22,** 1458–1468. ISSN: 1945-0494, 1093-4715 (Mar. 2017).

20. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System. en. *Metab. Eng.* **48,** 233–242. ISSN: 1096-7176, 1096-7184 (July 2018).

21. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908. ISSN: 1087-0156, 1546-1696 (Oct. 2017).

22. Lennen, R. M., Jensen, K., Mohammed, E. T., Malla, S., Borner, R. A., Chekina, K., Ozdemir, E., Bonde, I., Koza, A., Maury, J., Pedersen, L. E., Schoning, L. Y., Sonnenschein, N., Palsson, B. O., Sommer, M. O. A., Feist, A. M., Nielsen, A. T. & Herrgaard, M. J. *Adaptive laboratory evolution reveals general and specific chemical tolerance mechanisms and enhances biochemical production* en. May 2019.

23. Zhang, K., Woodruff, A. P., Xiong, M., Zhou, J. & Dhande, Y. K. A synthetic metabolic pathway for production of the platform chemical isobutyric acid. en. *ChemSusChem* **4,** 1068–1070. ISSN: 1864-5631, 1864-564X (Aug. 2011).

24. Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M. & Feist, A. M. Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. en. *Mol. Biol. Evol.* **31,** 2647–2662 (Oct. 2014).

25. Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. & Kim, J. F. Genome evolution and adaptation in a long-term experiment with Escherichia coli. en. *Nature* **461,** 1243–1247. ISSN: 0028-0836, 1476-4687 (Oct. 2009).

26. Blattner, F. R., Plunkett 3rd, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. The complete genome sequence of Escherichia coli K-12. en. *Science* **277,** 1453–1462. ISSN: 0036-8075 (Sept. 1997).

27. Wehrs, M., Tanjore, D., Eng, T., Lievense, J., Pray, T. R. & Mukhopadhyay, A. Engineering Robust Production Microbes for Large-Scale Cultivation. *Trends Microbiol.* **27,** 524–537. ISSN: 0966-842X (June 2019).

28. Lennen, R. M. & Herrgård, M. J. Combinatorial strategies for improving multiple-stress resistance in industrially relevant Escherichia coli strains. en. *Appl. Environ. Microbiol.* **80,** 6223–6242. ISSN: 0099-2240, 1098-5336 (Oct. 2014).

29. Qiao, W., Qiao, Y., Liu, F., Zhang, Y., Li, R., Wu, Z., Xu, H., Saris, P. E. J. & Qiao, M. Engineering Lactococcus lactis as a multi-stress tolerant biosynthetic chassis by deleting the prophage-related fragment. en. *Microb. Cell Fact.* **19,** 225. ISSN: 1475-2859 (Dec. 2020).

30. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. en. *Nucleic Acids Res.* **47,** D506–D515. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

31. Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I. & Karp, P. D. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. en. *Nucleic Acids Res.* **45,** D543–D550. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

32. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. The Pfam protein families database in 2019. en. *Nucleic Acids Res.* **47,** D427–D432. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

33. Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A. & Beltrao, P. A resource of variant effect predictions of single nucleotide variants in model organisms. en. *Mol. Syst. Biol.* **14,** e8430. ISSN: 1744-4292 (Dec. 2018).

34. Lamoureux, C. R., Choudhary, K. S., King, Z. A., Sandberg, T. E., Gao, Y., Sastry, A. V., Phaneuf, P. V., Choe, D., Cho, B.-K. & Palsson, B. O. The Bitome: digitized genomic features reveal fundamental genome organization. en. *Nucleic Acids Res.* **48,** 10157–10163. ISSN: 0305-1048, 1362-4962 (Oct. 2020).

Chapter 5

Conclusions

The recent advances in genome editing methods and computational analysis have made the whole living world's source code available for exploration, experimentation, and utilization. These advances have and promise to continue contributing to research and applications that benefit society. Thoroughly understanding organism-scale biological systems has become the next challenge to engineering biology. The natural process of adaptive evolution has been co-opted to accelerate advances in solving this challenge. The recent growth of adaptive evolution data promises new insights through meta-analysis. The work of this dissertation seeks to leverage an aggregated set of adaptive evolution mutational data to design strains with potentially value in applications.

The chapter "Consolidating Adaptive Laboratory Evolution Data into an Accessible and Computational Resource" described the implementation of a web-accessible platform, ALEdb, to integrate and report on adaptive evolution mutations. ALEdb provides for the need in the field of adaptive evolution for an accessible consolidation of ALE mutational data. Additionally, multiple features were implemented within ALEdb to enable intuitive navigation and analysis.

All ALEdb data can be exported for further custom analysis.

The chapter "Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity" described the development of analytical methods leveraging aggregated sources of genome annotations and ALEdb data to better identify and understand ALE key mutations. The identification of key mutations was enhanced by seeking mutations involved in statistically significant convergence events across multiple scales of genome annotations. The convergence of small mutated features onto broader functional annotations additionally enabled a top-down approach to mutation functional analysis, where one can first consider biological functions hosting mutations before investigating the numerous smaller mutated features. Mutated features across the aggregated ALE experiment data were statistically associated to experimental conditions. These associations provided evidence towards clarifying the selection pressures for mutated features among the multiple conditions describing each ALE experiment. Using this framework, it was found that (1) when mutated, the variety of non-coding regulatory features (promoters, transcription factor binding sites, terminators, etc.) had a higher probability of being involved in significant convergence events than genes, (2) the method of bottom-up convergence from small to large features on the multiscale annotation framework found more key mutations than when considering only genes and intergenic regions for mutation convergence, and (3) the selection pressures for case study key mutations were more clear due to strong statistical associations between mutated features and specific experimental conditions.

The chapter "Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data" described the use of ALE experiment meta-analysis results towards strain design. New analytical methods, as well as those developed in previous chapters, were used to identify and understand ALE mutation trends and their selection pressures. According to these

trends, novel variants with similar benefits to the original ALE mutations were designed and experimentally validated. These results have demonstrated how aggregated public ALE mutation data, rich annotations, and structural biology methods can provide enough evidence to interpret ALE mutation objectives towards strain design.

In this dissertation, resources and methods for ALE data-driven strain design were implemented and tested. A platform and underlying software pipeline were developed and deployed to consolidate ALE mutations and conditions data. A framework and methods for rich annotation and statistical analysis were developed to find mutational trends that serve to better identify and understand adaptive mutations than traditional means. Lastly, the information derived from the mutational trends was used to generate novel variants applicable to strain design. These results demonstrate how the large amount of available public ALE data can be aggregated and analyzed to design microbial strains.

Appendix A

Consolidating Adaptive Laboratory Evolution Data into an Accessible and Computational Resource - Supplementary Information

A.1 Methods

A.1.1 Mutation finding pipeline

Mutation data currently hosted on ALEdb are generated by the breseq mutation finding pipeline [1, 2]. Being that these samples come from different projects, various version of breseq were used in their mutation data generation. The sequencing reads used to generate the mutation data were subjected to quality control through either FastQC (`https:`

//www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the FastX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) or AfterQC [3].

A.1.2 Experimental evolution normalization for case study

In comparing endpoint mutations between different experiments, strategies are necessary to normalize experiments which have different replicate evolution counts and lengths. For very long evolution, such as the Long-term Experimental Evolution, samples at 2000 generations were considered endpoints. To normalize for differing amounts of replicate evolutions between experiments, the average proportion of each mutation type of interest is found across an experiment's replicates to represent the experiment. Additionally, no samples containing hypermutator strains were included in the case study described in the final section.

A.1.3 Supplementary Figures and Tables

A.2 Abbreviations

ALE: Adaptive laboratory evolution. COGs: Cluster of orthologous groups. DAG: Directed acyclic graph. TU: Transcription unit. TFBS: Transcription factor binding site. RBS: Ribosomal binding site. SNP: Single nucleotide polymorphism mutation. DEL: Deletion mutation. INS: Insertion mutation. MOB: Mobile element insertion mutation. CCR: Carbon catabolite repression. FBP: fructose-1,6-bisphosphate. EIIA: Enzyme II A. IIAGlc: Enzyme IIA (Glc). AC: Adenylate cyclase. cAMP: Cyclic adenosine monophosphate. PTS: Phosphotransferase system. $\Delta\Delta$Gpred: The difference between the free energy of unfolding the protein structure before and after a mutation. SIFT: Sorting intolerant from tolerant.

Table A.1: Test of independence between mutation type distributions of Figure 2.7A

mutation distributions	SBRG sample size	KEE sample size	LTEE sample size	two-tailed Mann-Whitney U statistic	two-tailed critical value α=0.05	significantly different?
SNP	11	12		60.5	33	no
SNP	11		12	68	33	no
SNP		12	12	63	37	no
SV	11	12		71	33	no
SV	11		12	63	33	no
SV		12	12	62	37	no
DEL	11		12	98	33	no
MOB	11		12	74	33	no
INS	11		12	65	33	no

Table A.2: Test of significant difference between a coding SNP distribution and given proportion of coding nucleotides within the bacterial genomes from Figure 2.7B

experimental sample	sample size	proportion of coding nucleotides within the bacterial genome	two-tailed Wilcoxon Signed Rank statistic	two-tailed critical value α=0.05	significantly different
SBRG	11	0.86	15	10	no
KEE	12	0.86	32	13	no
LTEE	12	0.86	15	13	no

A.3 Data Availability

ALEdb is freely available online at http://aledb.org and can be accessed with a JavaScript-enabled browser.

A.4 References

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

2. Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., Knoester, D. B., Reba, A. & Meyer, A. G. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. en. *BMC Genomics* **15,** 1039 (Nov. 2014).

3. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. & Gu, J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. en. *BMC Bioinformatics* **18,** 80 (Mar. 2017).

Appendix B

Causal mutations from adaptive laboratory evolution are outlined by multiple scales of genome annotations and condition-specificity - Supplementary Information

B.1 Methods

B.1.1 ALE experiments

This work focuses on ALE experiments with a starting strain of *E. coli* K-12 MG1655 or derivative thereof, due to the availability of experimental data and genome annotations. Thirty

five ALE experiments, each using a unique set of conditions and selection pressures, were exported from ALEdb [1] for this work [2–14]. These ALE experiments consist of 208 evolutions and 2641 mutations. Mutations in population samples with a frequency below 50% were filtered out to instead focus on mutations that demonstrate dominant selection within a sample. When inspecting experiments for key mutations, endpoint samples containing known hypermutator strains were discarded in an effort to focus on less complex genotypes.

B.1.2 ALE experiment mutation organization

Only mutations from endpoint flask samples (i.e., the flask that contained the final sequenced sample) were considered for the ALE experiments used in this work. Both clonal and population samples were included in this work. If the same mutation was available in more than one sample from an endpoint flask, the instance of the mutation with the highest sample frequency was chosen to represent all instances of the mutation from the flask.

B.1.3 ALE experiment conditions

The ALE experiment conditions metadata used in the associations was gathered from the metadata reports available from ALEdb [1] for each ALE experiment considered in this work.

B.1.4 Software for analysis and figure generation

Quantitative plots

Unless otherwise stated, figure plots were generated using Matplotlib version 3.0.3 [15] and Seaborn version 0.9 Python software packages [16].

Flow diagrams

The mutation flow diagrams of Figure 3.1c, Figure 3.3a, Figure 3.4a, Figure 3.5a and Supplemental Figure 3.1 were generated using the Floweaver Python software package [17].

Mutation needle plot diagram

The mutation needle plot of Figure 3.3e was generated using the muts-needle-plot Javascript software package [18].

B.1.5 Mutation calling

The breseq pipeline version 0.33.1 [19] was used to map the DNA-seq reads to an *E. coli* K-12 MG1655 reference genome. The reference genome was either the NCBI accession NC_000913 version 3 reference genome or a derivative based on this NCBI genome [20]. DNA-seq quality control was accomplished using the software AfterQC version 0.9.7 [21].

B.1.6 Mutation effect prediction

The predicted disrupted effects of mutations to the genes in the glycerol evolution results were generated using the mutfunc web application [22].

B.1.7 Software scripts

Software scripts and data essential to the publication is available through a GitHub repository [23].

B.1.8 Supplementary Figures and Tables

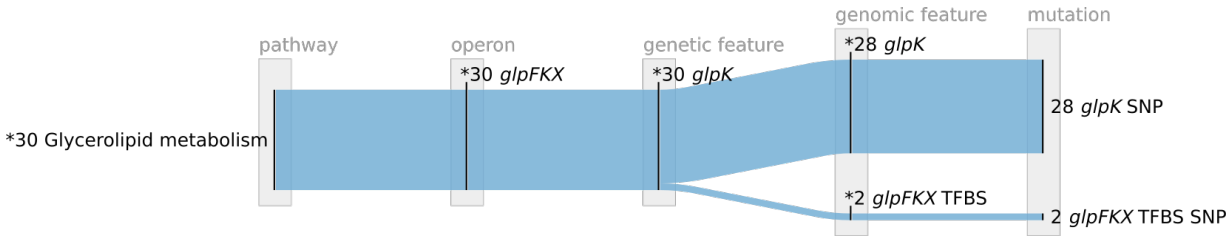


Figure B.1: The mutation convergence and significance of features associated with *glpK* and glycerol metabolism pathway.

Table B.2: Table of *glpK* SNP details.

AA position	AA change	summary
17	S17A	uncharacterized
34	R34H	uncharacterized
34	R34H	uncharacterized
54	W54L	uncharacterized
55	A55T	uncharacterized
56	T56A	uncharacterized
59	S59Y	subunit interaction interface [24]
59	S59Y	subunit interaction interface [24]
59	S59Y	subunit interaction interface [24]
59	S59Y	subunit interaction interface [24]
66	A66S	associated to FBP binding [25]
73	D73V	subunit interaction interface [26]
73	D73V	subunit interaction interface [26]
73	D73V	subunit interaction interface [26]

Table B.2: Table of *glpK* SNP details, continued.

AA position	AA change	summary
73	D73V	subunit interaction interface [26]
73	D73V	subunit interaction interface [26]
73	D73V	subunit interaction interface [26]
96	K96E	uncharacterized
96	K96Q	uncharacterized
96	K96Q	uncharacterized
96	K96Q	uncharacterized
97	P97H	Computationally predicted stability [22]
173	K173T	Computationally predicted stability [22]
229	N229H	uncharacterized
229	N229K	uncharacterized
236	T236K	FBP binding site [27]
236	T236K	FBP binding site [27]
441	A441V	uncharacterized

B.2 Availability of data and materials

The datasets supporting the conclusions of this article are available in the GitHub repository [23], ALEdb database [1], and within this article's supplementary tables.

Table B.1: Table of *crr* mutation details

start AA position	stop AA position	mut type	details	ΔΔGpred (destabilizing)	SIFT score (deleterious)
34	34	SNP	I34T	3.47091	0.0074979
39	39	SNP	D39H	2.49214	0
44	44	SNP	E44V	NA	0
48	48	SNP	G48V	18.6249	0
50	50	SNP	G50S	13.6354	0
72	72	SNP	F72V	NA	0
72	72	SNP	F72V	NA	0
93	93	SNP	G93S	7.2727	0
138	138	SNP	P138Q	3.46287	0
138	138	SNP	P138Q	3.46287	0

B.3 References

1. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* **47,** D1164–D1171. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

2. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Appl. Environ. Microbiol.* **81,** 17–30 (Jan. 2015).

3. Charusanti, P., Conrad, T. M., Knight, E. M., Venkataraman, K., Fong, N. L., Xie, B., Gao, Y. & Palsson, B. O. Genetic Basis of Growth Adaptation of Escherichia coli after Deletion of pgi, a Major Metabolic Gene. *PLoS Genet.* **6,** e1001186 (Nov. 2010).

4. LaCroix, R. A., Palsson, B. O. & Feist, A. M. A Model for Designing Adaptive Laboratory Evolution Experiments. en. *Appl. Environ. Microbiol.* **83** (Apr. 2017).

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

6. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System. en. *Metab. Eng.* **48,** 233–242. ISSN: 1096-7176, 1096-7184 (July 2018).

7. Sandberg, T. E., Pedersen, M., LaCroix, R. A., Ebrahim, A., Bonde, M., Herrgard, M. J., Palsson, B. O., Sommer, M. & Feist, A. M. Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. en. *Mol. Biol. Evol.* **31,** 2647–2662 (Oct. 2014).

8. Sandberg, T. E., Long, C. P., Gonzalez, J. E., Feist, A. M., *et al.* Evolution of E. coli on [U-13C] glucose reveals a negligible isotopic influence on metabolism and physiology. *PLoS* (2016).

9. Sandberg, T. E., Lloyd, C. J., Palsson, B. O. & Feist, A. M. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. en. *Appl. Environ. Microbiol.* **83** (July 2017).

10. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Multiple Optimal Phenotypes Overcome Redox and Glycolytic Intermediate Metabolite Imbalances in Escherichia coli pgi Knockout Evolutions. en. *Appl. Environ. Microbiol.* **84.** ISSN: 0099-2240, 1098-5336 (Oct. 2018).

11. McCloskey, D., Xu, S., Sandberg, T. E., Brunk, E., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptation to the coupling of glycolysis to toxic methylglyoxal production in

tpiA deletion strains of Escherichia coli requires synchronized and counterintuitive genetic changes. en. *Metab. Eng.* **48,** 82–93. ISSN: 1096-7176, 1096-7184 (July 2018).

12. Anand, A., Chen, K., Catoiu, E., Sastry, A. V., Olson, C. A., Sandberg, T. E., Seif, Y., Xu, S., Szubin, R., Yang, L., Feist, A. M. & Palsson, B. O. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. en. *Mol. Biol. Evol.* ISSN: 0737-4038, 1537-1719 (Oct. 2019).

13. Lennen, R. M., Jensen, K., Mohammed, E. T., Malla, S., Borner, R. A., Chekina, K., Ozdemir, E., Bonde, I., Koza, A., Maury, J., Pedersen, L. E., Schoning, L. Y., Sonnenschein, N., Palsson, B. O., Sommer, M. O. A., Feist, A. M., Nielsen, A. T. & Herrgaard, M. J. *Adaptive laboratory evolution reveals general and specific chemical tolerance mechanisms and enhances biochemical production* en. May 2019.

14. Guzman, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, A., Papp, H., de Raad, M., King, Z. A., Hefner, Y., Northen, T. R., Notebaart, R. A., Pal, C., Palsson, B. O., Papp, B. & Feist, A. M. Enzyme promiscuity shapes adaptation to novel growth substrates. en. *Mol. Syst. Biol.* **15,** e8462. ISSN: 1744-4292 (Apr. 2019).

15. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9,** 90–95. ISSN: 1521-9615 (May 2007).

16. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Brunner, T., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian & Qalieh, A. *mwaskom/seaborn: v0.9.0 (July 2018)* July 2018.

17. Lupton, R. C. & Allwood, J. M. Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. *Resour. Conserv. Recycl.* **124,** 141–151. ISSN: 0921-3449 (Sept. 2017).

18. Schroeder, M. P. & Lopez-Bigas, N. *muts-needle-plot: Mutations Needle Plot v0.8.0* Jan. 2015.

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

20. Phaneuf, P. *SBRG/bop27refseq: Zenodo DOI release (Version v1.4.1)* Dec. 2019.

21. Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. & Gu, J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. en. *BMC Bioinformatics* **18,** 80 (Mar. 2017).

22. Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A. & Beltrao, P. A resource of variant effect predictions of single nucleotide variants in model organisms. en. *Mol. Syst. Biol.* **14,** e8430. ISSN: 1744-4292 (Dec. 2018).

23. Phaneuf, P. *ALE Variant Analysis scripts and data v0.1.0* Apr. 2020.

24. Bystrom, C. E., Pettigrew, D. W., Branchaud, B. P., O'Brien, P. & Remington, S. J. Crystal structures of Escherichia coli glycerol kinase variant S58–¿W in complex with nonhydrolyzable ATP analogues reveal a putative active conformation of the enzyme as a result of domain motion. en. *Biochemistry* **38,** 3508–3518. ISSN: 0006-2960 (Mar. 1999).

25. Feese, M. D., Faber, H. R., Bystrom, C. E., Pettigrew, D. W. & Remington, S. J. Glycerol kinase from Escherichia coli and an Ala65→Thr mutant: the crystal structures reveal conformational changes with implications for allosteric regulation. *Structure* **6,** 1407–1418. ISSN: 0969-2126 (Nov. 1998).

26. Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R. & Palsson, B. O. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. en. *Nat. Genet.* **38,** 1406–1412. ISSN: 1061-4036 (Dec. 2006).

27. UniProt Consortium, T. UniProt: the universal protein knowledgebase. en. *Nucleic Acids Res.* **46,** 2699. ISSN: 0305-1048, 1362-4962 (Mar. 2018).

Appendix C

Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data - Supplementary Information

C.1 Methods

C.1.1 Biological material

Mutant strains were generated following a CRISPR/Cas9-assisted protocol outlined by Zhao *et al.* [1] (Table **??**). This method relies on Cas9 to cut the genome of the starting strain while leaving intact the successfully mutated genome. A single plasmid encoding the CRISPR/Cas9 and Lambda Red Recombinase systems along with repair arms and a 20 nt guide

RNA targeting the starting strain sequence was constructed using Golden Gate Assembly. In this case the repair arms were generated using PCR instead of annealing two oligos. Depending on the target mutation, one of three strategies for the placement of the 20 nt guide RNA were used: 1) next to a PAM sequence if that PAM sequence was eliminated by the mutation 2) spanning the bases targeted for mutation if that region was next to a PAM sequence or 3) close to the region being mutated and next to a PAM sequence that could be eliminated by introducing a synonymous codon with the repair arms. When option 2 was used to introduce a single nucleotide change, one of the 20 bases of the guide RNA unrelated to that position was deliberately changed so that the successfully mutated strain would have two mismatched bases with respect to the guide RNA while the starting strain would have just one mismatch. Colonies were screened using ARMS PCR in which one of the primers was designed to work on the starting strain and not on the mutated sequence. All mutations were verified by Sanger sequencing an amplicon generated with primers targeting the genome distal to the ends of the repair arms to avoid sequencing the plasmid. Finally the plasmid was eliminated by growth at 37°C, verified by parallel plating on media with and without Kanamycin.

Each chromosomal *cyaA*, *pgi* and *pykF* deletion was introduced using a temperature-sensitive pGE3 carrying lambda-recombinases and MAD7 nuclease (MADzymeTM) which sequence obtained from Inscripta Inc. Briefly, *E. coli* MG1655 was transformed with a temperature-sensitive pGE3 (Table C.1, Table C.2), modified from pMP11 by replacing Cas9 into MAD7 [2]. Lambda recombinases were induced with 0.2% arabinose at 37℃ for 45 min. Then, induced cells were transformed with 200 ng gRNA plasmid together with 100 pmol synthetic oligo containing each flanking 45 bp homologous sequence for each gene. Upon transformation, cells were recovered at 30℃ for 1 hr 30 min. Then, the cells were transferred to 2 ml of LB containing ampicillin

Table C.1: Table of plasmids for gene knockouts.

Plasmids	Genotype	Comments
pGE3	araC Para::gam-bet-exo tetR PlacI::tetO::gRNA of pgRNA Pj23105::MAD7 Am R SC101(ts)	[2]
pgRNA-bacteria	Pj23119::gRNA CmR BR322	[2]
pgRNA-*cyaA*	Pj23119::gRNA_cyaA CmR BR322	21bp gRNA sequence for *cyaA*
pgRNA-*pgi*	Pj23119::gRNA_pgi CmR BR322	21bp gRNA sequence for *pgi*
pgRNA-*pykF*	Pj23119::gRNA_pykF CmR BR322	21bp gRNA sequence for *pykF*

and chloramphenicol and grew overnight at 30℃. Knockout transformants were isolated by plating and validated by colony PCR and Sanger sequencing. Loss of guide RNA plasmid was done by growing confirmed isolates in LB containing ampicillin and 200 ug/L of anhydrotetracycline at 30℃ for at least 6 hours. Loss of pGE3 was achieved by propagating at 37℃. Loss of both plasmids was validated by checking the cell's sensitivity to antibiotics. *cyaA* knockout was performed on Δ*pgi* strain for construction of *pgi* and *cyaA* double knockout strain.

Table C.2: Table of primers for gene knockouts.

Primer	Sequence
cyaA_MAGE	GATAAGCCTCGCTTTCCGGCACGTTCATCACGAAAAATATTGCTGCTC AATATAGAGGTACAAGACGTATCGCCTGATTTGCTACCCGTC
pgi_MAGE	CTTATCCGGCCTACATATCGACGATGATTAACCGCGCCACGCTTTTGG ATTGATGTTTTTCATTAGCAATACTCTTCTGATTTTGAG
pykF_MAGE	TTCTCAACTTAAAGACTAAGACTGTCATGAAAAAGACCAAAATTTCTG TTCACGTCCTGTAATATTGCTTTTGTGAATTAATTTGT
cyaA_Flank_F	TTGTGGTGAAACGTAGACGCCTG
cyaA_Flank_R	CGGCGTACTGACCATTACCTTTG
pgi_Flank_F	AATGCTTCACTGCGCTAAG
pgi_Flank_R	TACATATCGGCATCGACCTG
pykF_Flank_F	AATTTCAGCGTATAATGCGCGCC
pykF_Flank_R	GGTCAACTTTAGCGTTCAGAGTCTGAA

Table C.3: Golden Gate CRISPR Oligos.

New Strain Designation	Mutation Description	Mutation	Does Mutation Remove PAM motif?	Does Mutation Significantly Change Protspacer	Is the mutation within a gRNA region?	Intermidiary synonymous codon needed?	Synonymous Codon	Resulting Sequence	Guide Sense Oligo AGCG	Guide Antisense Oligo AAAC	DONOR F GTGC	DONOR R GAGC	Left Donor Arm F GTGC	Left Donor Arm R	Right Donor Arm F	Right Donor Arm R GAGC
PVP002	4,117,046 (2 bp) GG→CC	S59W TCC -> TGG	Yes					AATGGAA ATCTGGG CCACCCA AAGCTGG ACGCTGG TAGAAGT GCTGGCG A	AGCGCGCC AGCACTTCT ACCAGCG	AAACCGCTGG TAGAAGTGCT GGCG						
PVP003	4,117,047 G→A	S59F TCC -> TTC	Yes					AATGGAA ATCTGGG CCACCCA AAGCTTC ACGCTGG TAGAAGT GCTGGCG A	AGCGCGCC AGCACTTCT ACCAGCG	AAACCGCTGG TAGAAGTGCT GGCG						
PVP005	cyaA with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 540			Yes	Yes			CCCGCTG CGCTTAC CTGCACC GACATAAT AGTGACC GAAGGCG CTCTACA GCCCGTG TG	AGCGCTTAC CTGCACCG ACACCGA	AAACTCGGTG TCGGTGCAG GTAAG			CCAGGTCTCAGT GCGGGCTAACCG TTCAGGTTGGT	CCAGGTCTCAAT TATGTCGGTGCA GGTAAGC	CCAGGTCTCATAA TAGTGACCGAAG GCGCTCTACAG	CCAGGTCTCAGAG CCGGCGTCCTGAT GCAT
PVP007	cyaA with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 540			Yes	Yes			CCCGCTG CGCTTAC CTGCACC GACATAAT AGTGACC GAAGGCG CTCTACA GCCCGTG TG	AGCGCTTAC CTGCACCG ACACCGA	AAACTCGGTG TCGGTGCAG GTAAG	CCAGGTCT CAGTGCGG GCTAACCG TTCAGGTT GGT	CCAGGTCT CAGAGCC GGCGTCCT GATGCAT	CCAGGTCTCAGT GCGGGCTAACCG TTCAGGTTGGT	CCAGGTCTCAAT TATGTCGGTGCA GGTAAGC	CCAGGTCTCATAA TAGTGACCGAAG GCGCTCTACAG	CCAGGTCTCAGAG CCGGCGTCCTGAT GCAT
PVP010	pykF with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 253			Yes	Yes			TGGCGAC CTGGGTG TAGAAATC CCGTAATA GTGAGTA GAAGAAG TTATCTTC GCCCAGA	AGCGCGAA GATAACTTC TTCTACC	AAACGGTAGA AGAAGTTATC TTCG			CCAGGTCTCAGT GCCCTGGCGTTT CCATTGCT	CCAGGTCTCAAT TACGGGATTTCT ACACCCAGGT	CCAGGTCTCATAA TAGTGAGTAGAAG AAGTTATCTTCGC CCAG	CCAGGTCTCAGAG CGCGCAGTTTACG GTTGTCATT
PVP006	cyaA Δ5 bp (1365-1369/2547 nt) GCCGG															
PVP001	4,117,047 G→T	S59Y TCC -> TAC	Yes					AATGGAA ATCTGGG CCACCCA AAGCTAC ACGCTGG TAGAAGT GCTGGCG A	AGCGCGCC AGCACTTCT ACCAGCG	AAACCGCTGG TAGAAGTGCT GGCG	CCAGGTCT CAGTGCGC GCGATTGT AGGTGCAT T	CCAGGTCT CAGAGCCT CGCGAGA GCCTTCC				
PVP004	cyaA Δ5 bp (1365-1369/2547 nt) GCCGG			YES	Yes	Yes	S445 TCG to TCC	CGAATCT GACCTTTA TTTATGTG CC|GCCG GGCTAAC CGTTCAG GTTGGTA	AGCGTATTT ATGTGCCG CCGGGCC	AAACGGCCC GGCGGCACAT AAATA	CCAGGTCT CAGTGCCG CAGTGTTG AGCCAGTT	CCAGGTCT CAGAGCGT AAGCGCA GCGGGAA A				

C.1.2 Physiological characterizations

The growth rates of strain clones were screened by inoculating cells from an overnight culture to a low OD and then sampling the OD600 until the stationary phase was reached. A linear regression of the log-linear region was computed using the *linregress* function from the *scipy.stats* Python package and the growth rates were determined from the resulting slopes. Screen condition details are given in Table C.4.

Table C.4: Strain design physiological characterizations details. Media was M9 in all cases. Temperature was 37°C in all cases except the glycerol GlpK mutant and glycerol CyaA mutant assays, which used a temperature of 30°C. Supplement was of NaCl(0.5g/L) except with the isobutyrate toxic supplement assay, where the supplement was NaCl(0.5g/L), isobutyrate(12.5g/L). Nitrogen source as NH4Cl(1g/L) in all cases. Phosphorous source was KH2PO4(3g/L) and Na2HPO4(6.8g/L) in all cases. Sulfur source was MgSO4(0.24g/L) in all cases. Calcium source was CaCl2(0.1g/L) in all cases.

assay description	strain	carbon source
glycerol GlpK mutant	BOP27: WT	glycerol (2.5 g/L)
glycerol GlpK mutant	PVP001: *glpK* S59Y (TCC→TAC)	glycerol (2.5 g/L)
glycerol GlpK mutant	PVP002: *glpK* S59W (TCC→TGG)	glycerol (2.5 g/L)
glycerol GlpK mutant	PVP003: *glpK* S59F (TCC→TTC)	glycerol (2.5 g/L)
glycerol CyaA mutant	BOP27: WT	glycerol (2.5 g/L)
glycerol CyaA mutant	Δ*cyaA*	glycerol (2.5 g/L)
glycerol CyaA mutant	PVP004: *cyaA* Δ5 bp (1365-1369/2547 nt)	glycerol (2.5 g/L)

Table C.4: Strain design physiological characterizations details, continued.

assay description	strain	carbon source
glycerol CyaA mutant	PVP005: *cyaA* with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 540	glycerol (2.5 g/L)
gene perturbation Δpgi	Δpgi	glucose (4 g/L)
gene perturbation Δpgi	Δpgi $\Delta cyaA$	glucose (4 g/L)
gene perturbation Δpgi	PVP006: Δpgi *cyaA* Δ5 BP (1365-1369/2547 NT, 455-456/848 AA)	glucose (4 g/L)
gene perturbation Δpgi	PVP007: Δpgi *cyaA* with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 540	glucose (4 g/L)
isobutyrate toxic supplement	BOP27: WT	glucose (4 g/L)
isobutyrate toxic supplement	ΔpykF	glucose (4 g/L)

Table C.4: Strain design physiological characterizations details, continued.

assay description	strain	carbon source
isobutyrate toxic supplement	PVP009: pykF with 1x adenine (A) inserted at nucleotide 798	glucose (4 g/L)
isobutyrate toxic supplement	PVP010: pykF with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 253	glucose (4 g/L)
glucose selection pressure check	BOP27: WT	glycerol (2.5 g/L)
glucose selection pressure check	ΔpykF	glycerol (2.5 g/L)
glucose selection pressure check	PVP009: pykF with 1x adenine (A) inserted at nucleotide 798	glycerol (2.5 g/L)
glucose selection pressure check	PVP010: pykF with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 253	glycerol (2.5 g/L)
high glucose concentration	BOP27: WT	glucose (4 g/L)
high glucose concentration	ΔpykF	glucose (4 g/L)

Table C.4: Strain design physiological characterizations details, continued.

assay description	strain	carbon source
high glucose concentration	PVP009: pykF with 1x adenine (A) inserted at nucleotide 798	glucose (4 g/L)
high glucose concentration	PVP010: pykF with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 253	glucose (4 g/L)
low glucose concentration	BOP27: WT	glucose (0.200 g/L)
low glucose concentration	ΔpykF	glucose (0.200 g/L)
low glucose concentration	PVP009: pykF with 1x adenine (A) inserted at nucleotide 798	glucose (0.200 g/L)
low glucose concentration	PVP010: pykF with 3x stop codon sequence inserted (e.g., TAATAGTGA) @ AA 253	glucose (0.200 g/L)

C.1.3 Residue properties

Residue properties were acquired from the software package *ssbio* [3].

C.1.4 SIFT and ΔΔG scores for the mutation effect prediction of amino acid substitutions

In the case studies, two more effect types were predicted for mutations: the disruption of a potential function and of structural stability. Both of these were specific to coding regions. Functional disruptions were assumed according to significant SIFT scores (SIFT score < 0.05). Structural disruptions were assumed according to significant ΔΔG scores ($\Delta\Delta G > 2$). SIFT and ΔΔG scores were acquired from Mutfunc [4].

C.1.5 Gene product feature annotations

GlpK features were acquired from Uniprot [5], EcoCyc [6], Pfam [7], and Mutfunc [4] resources.

C.1.6 Gene product structures and distances between mutated residues and features

Gene product structures and distances between mutated residues and features Distances between mutated residues and the residues of functional features were calculated using gene product 3D structures and the Cartesian distance formula. GlpK was represented by the 3EZW PDB model. Structures for Crr, CyaA, and PykF were obtained from those provided with the iML1515 model of E. coli K-12 MG1655 [8].

C.1.7 Software scripts

Software scripts and data essential to the publication are available through 2 github repositories: https://github.com/Aletechdev/ava and https://github.com/Aletechdev/gdmuts.

C.1.8 Availability of data and materials

The datasets supporting the conclusions of this article are available in the ALEdb database [9] and the github repositories https://github.com/Aletechdev/ava and https://github.com/Aletechdev/gdmuts.

C.1.9 Mutation data cleaning

The mutations from ALEdb are initially described relative to the sample and the genome reference. Since some of these experiments include midpoint samples, mutations that emerge within a midpoint sample and carried through to an endpoint sample can get counted more than once per ALE, which would result in an inappropriately inflated mutation count. Unique ALE mutations were therefore only considered once per ALE. Starting strain mutations were filtered out of the ALE experiment mutation datasets according to their publications after being exported from ALEdb. When identified, samples containing hypermutators were also removed from the dataset. Mutations in population samples with a frequency below 50% were filtered out to instead focus on mutations that demonstrate dominant selection within a sample. In calculating correlations between mutated genomic features and generating the network diagrams of multi-scaled mutated features, large deletions were removed to filter out large sets of mutated features that were only mutated once.

C.1.10 Quantitative plots

Unless otherwise stated, figure plots were generated using Matplotlib version 3.0.3 [10] and Seaborn version 0.11 [11] Python software packages.

C.1.11 Network diagrams

The network diagrams of multiscale mutated features were generated using Cytoscape.js [12].

C.1.12 Mutation Needle Plots

The mutation needle plots were generated using the trackViewer R software package [13].

C.1.13 Oncoplots

The oncoplots were generated using the ComplexHeatmap R software package [14].

C.1.14 3D protein structures

The visualizations for the 3D protein structures were generated using the NGL software package [15].

C.1.15 Supplementary Figures and Tables

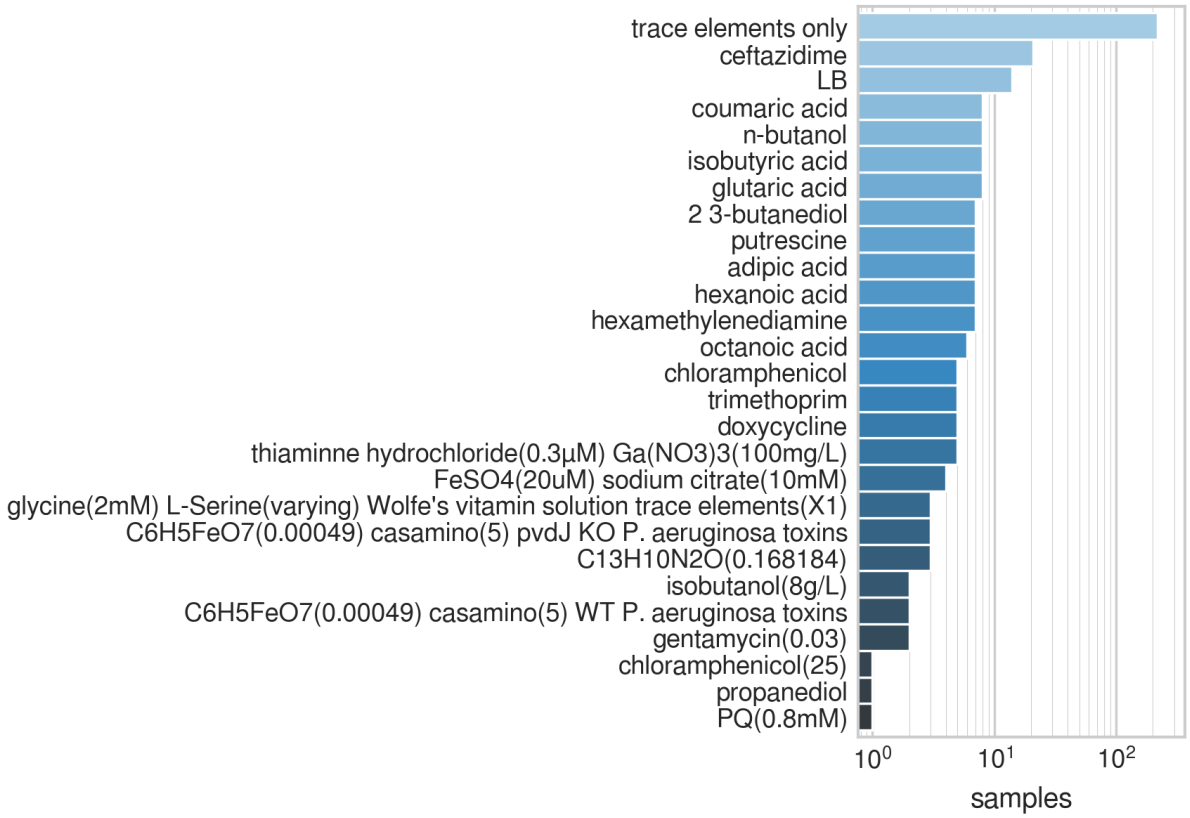


Figure C.1: The different conditions for the supplement condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

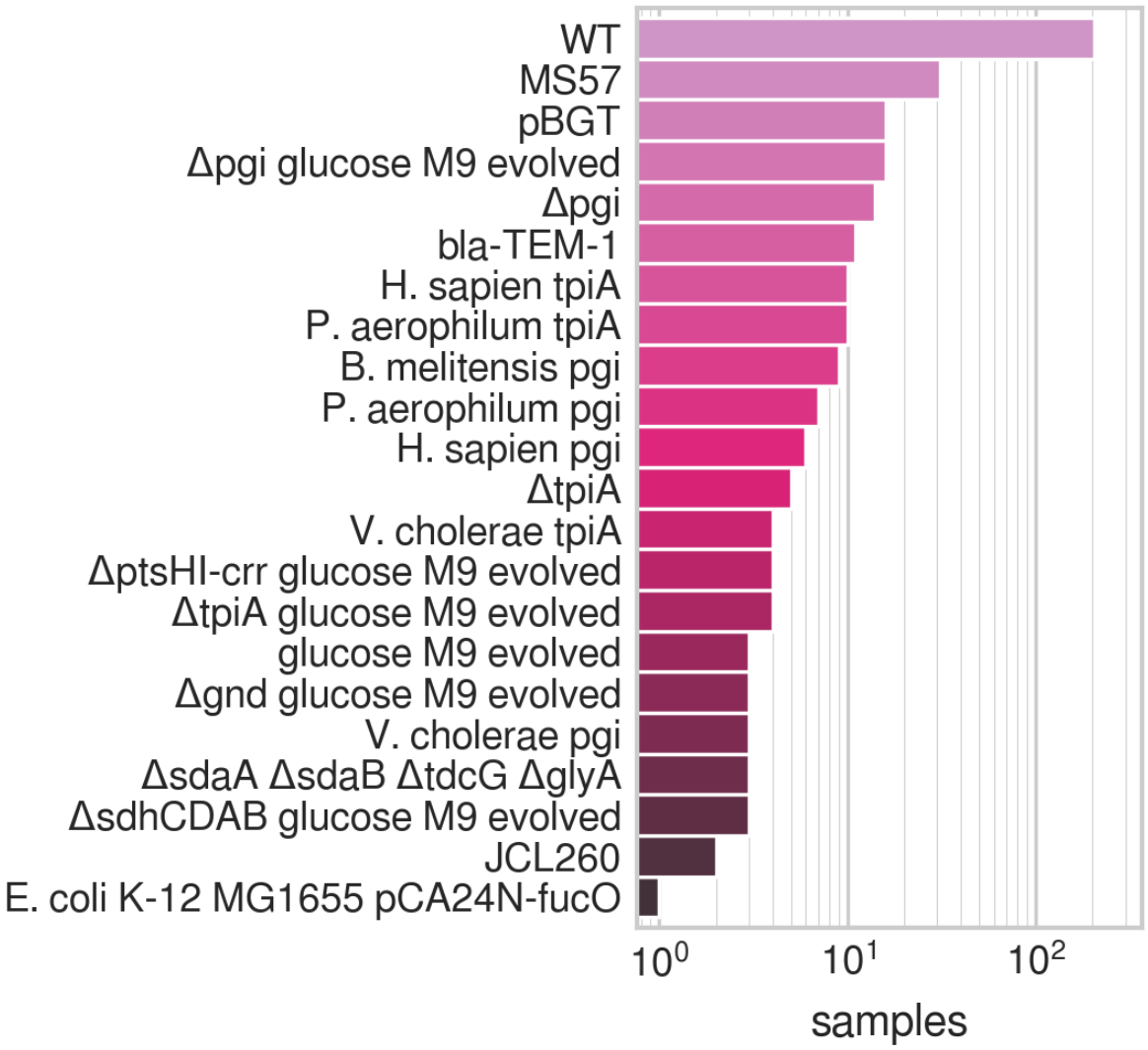


Figure C.2: The different conditions for the strain condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

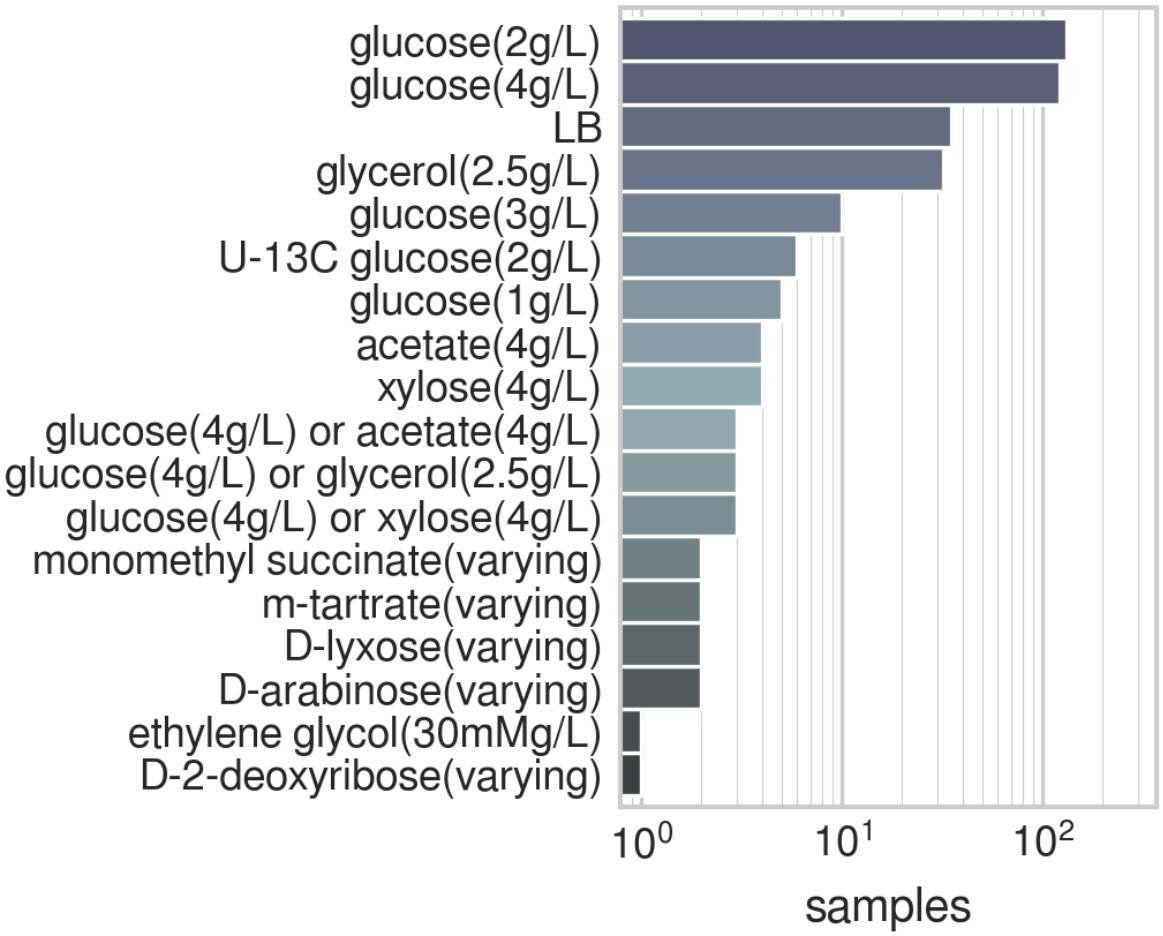


Figure C.3: The different conditions for the carbon source condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

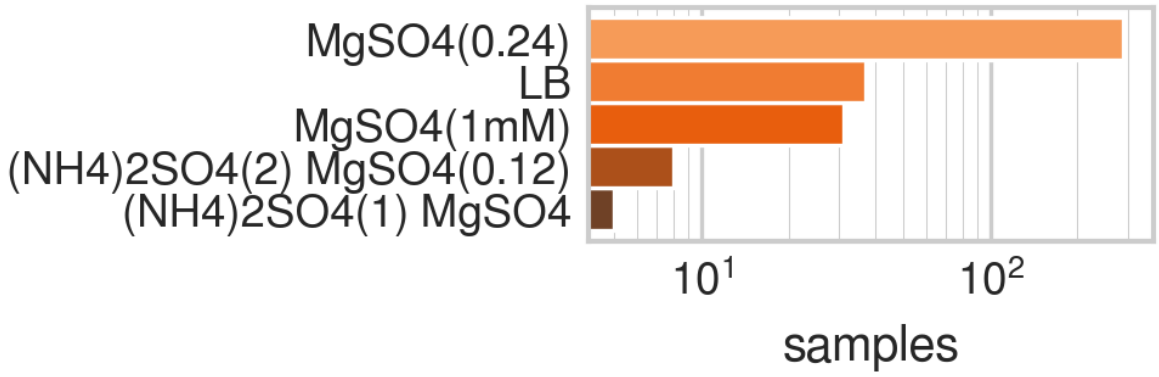


Figure C.4: The different conditions for the sulfur source condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

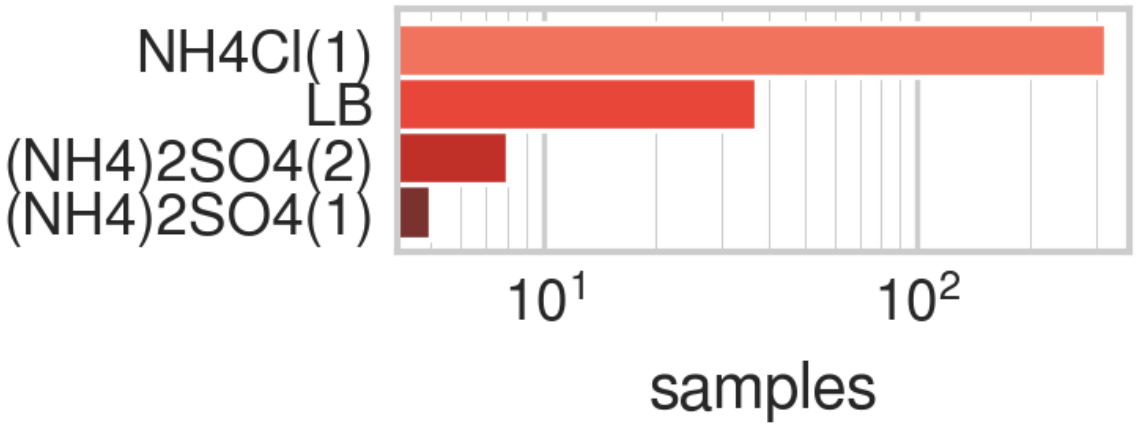


Figure C.5: The different conditions for the nitrogen source condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

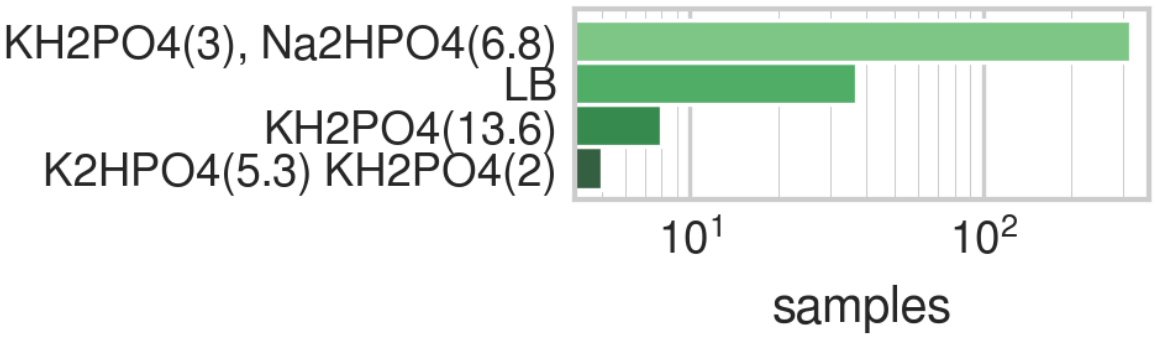


Figure C.6: The different conditions for the phosphorous source condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

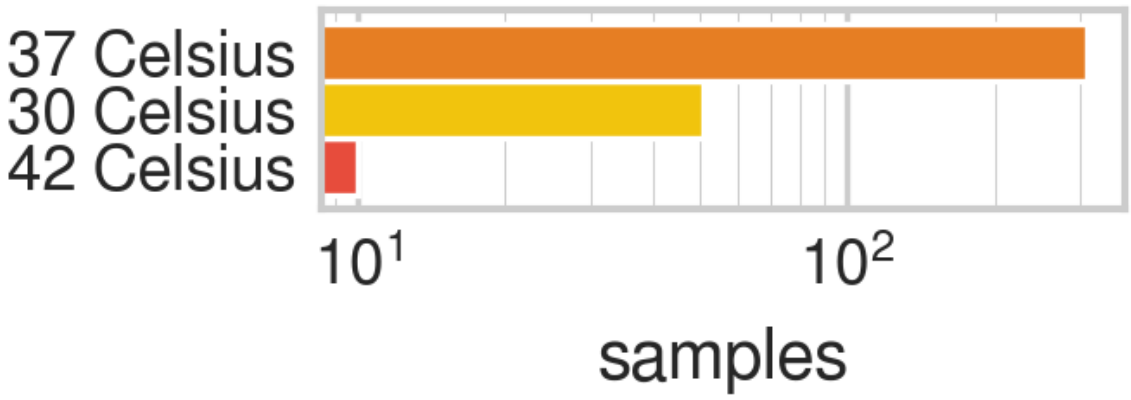


Figure C.7: The different conditions for the temperature condition category.

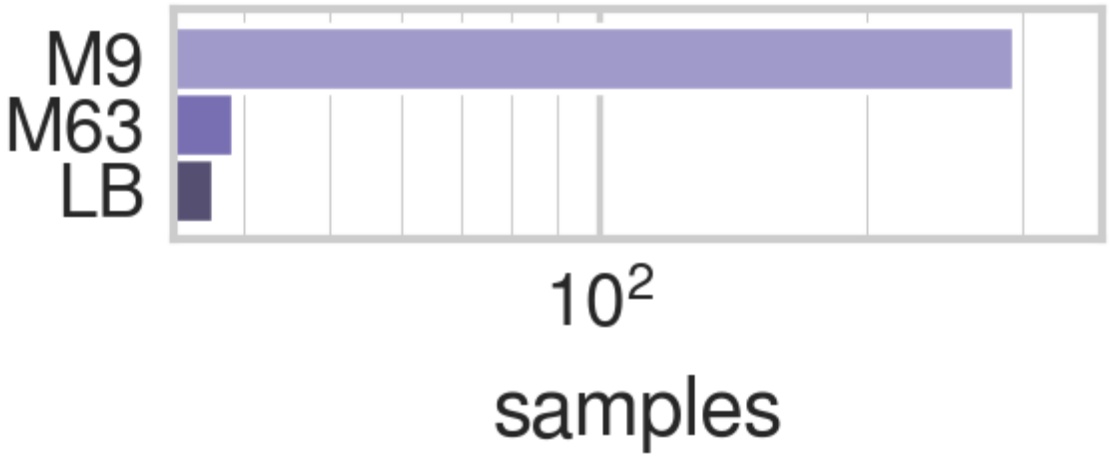


Figure C.8: The different conditions for the base-media condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

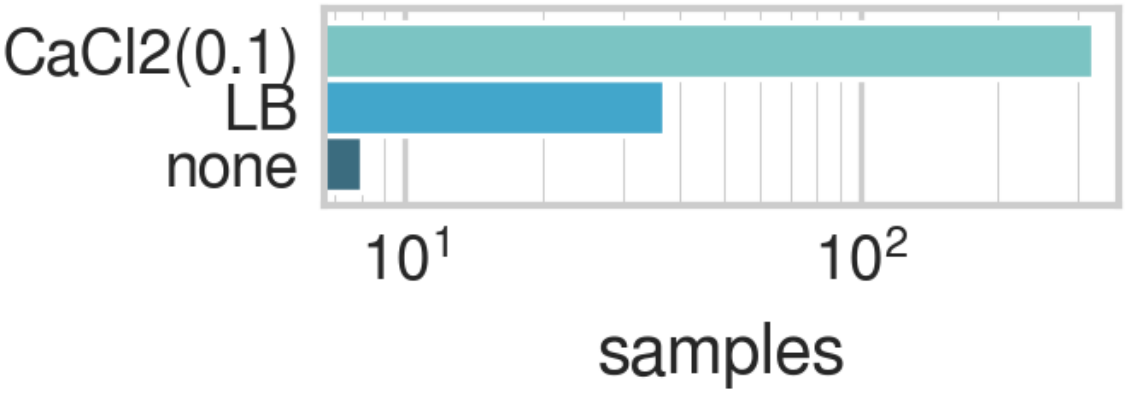


Figure C.9: The different conditions for the calcium source condition category. Values within parentheses represent concentrations in g/L unless otherwise stated.

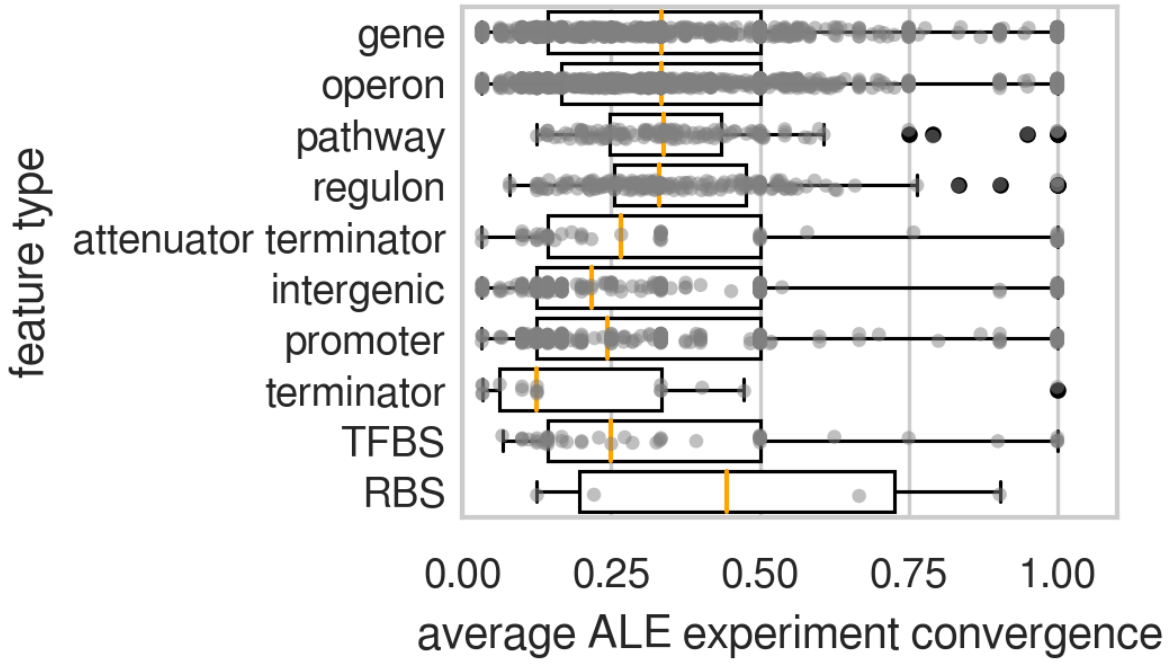


Figure C.10: The average proportion of ALE experiment replicates that a feature was mutated (convergence) for all features and ALE experiments.

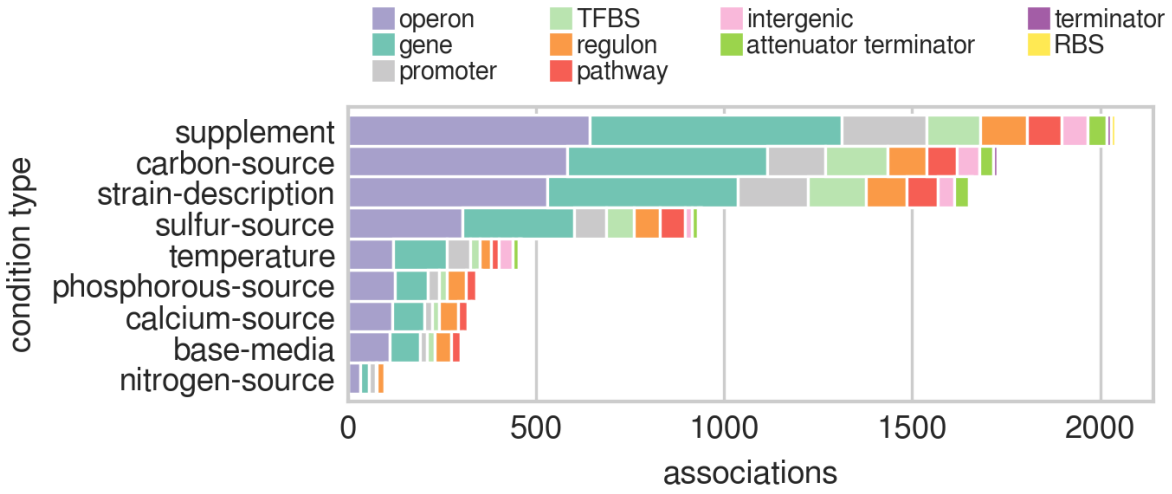


Figure C.11: The total amount of statistically significant associations between unique features and conditions according to condition types.

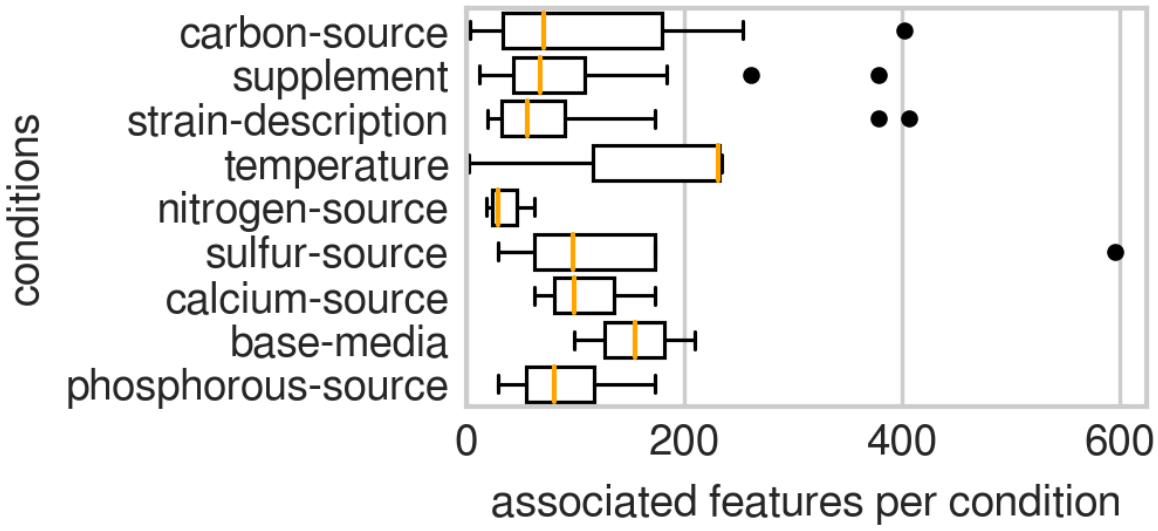


Figure C.12: The distribution of associated unique features per condition.

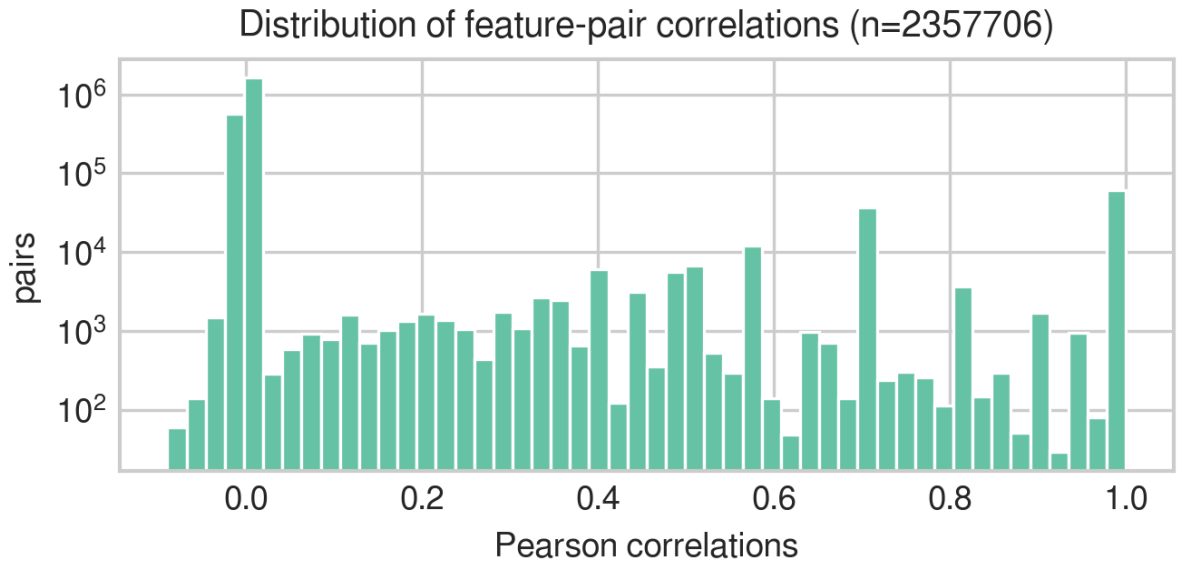


Figure C.13: The distribution of Pearson correlation coefficients for all genomic feature pairs.

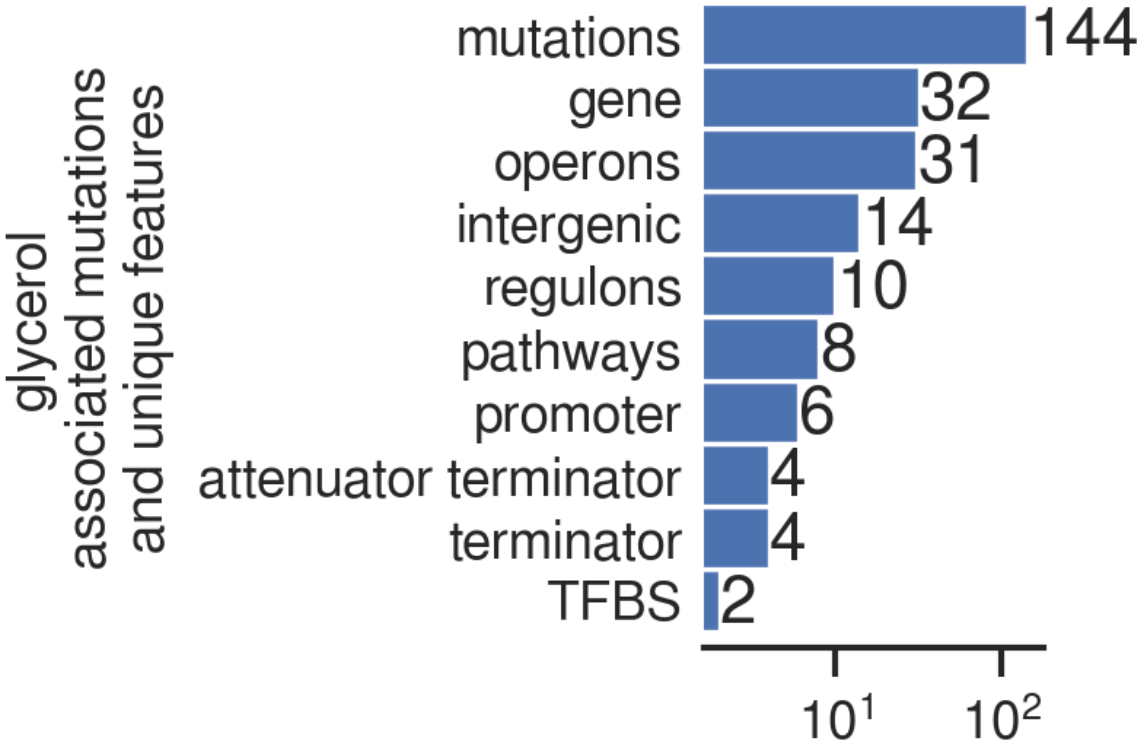


Figure C.14: The number of mutations and unique mutated features significantly associated with glycerol as a carbon source.

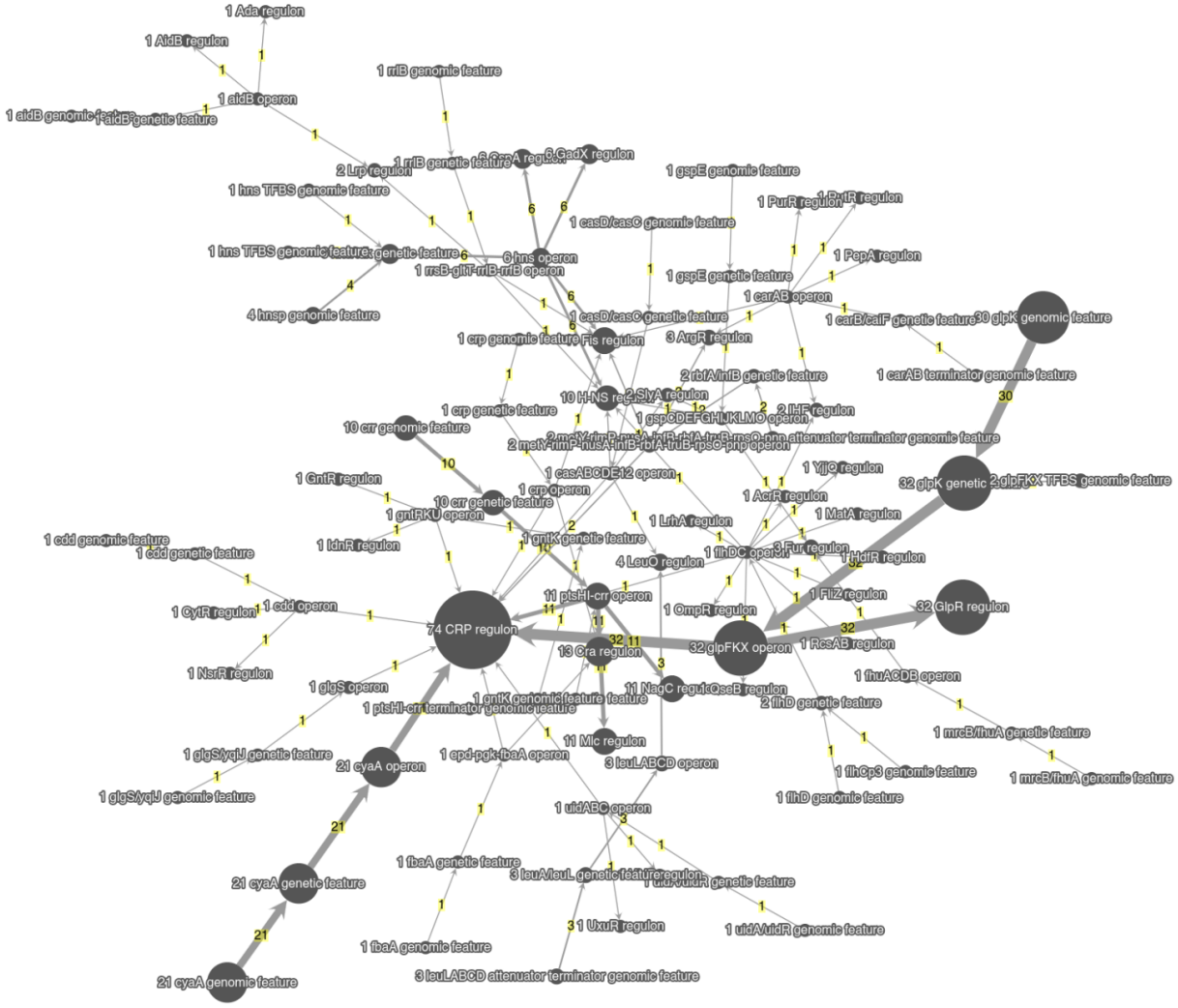


Figure C.15: A sub-network of mutated features associated with the glycerol as a carbon source and includes the CRP regulon demonstrates mutations to *glpK*, *cyaA*, and *crr* being the primary contributors to mutation convergence on the CRP regulon. The nodes of the network represent features annotated on the genome and the edges represent connections between features. The weight and directionality of an edge represents the amount of smaller mutated features that can be represented by the larger feature. The weight of a node represents the amount of mutated features that converge on the specific genome annotation.

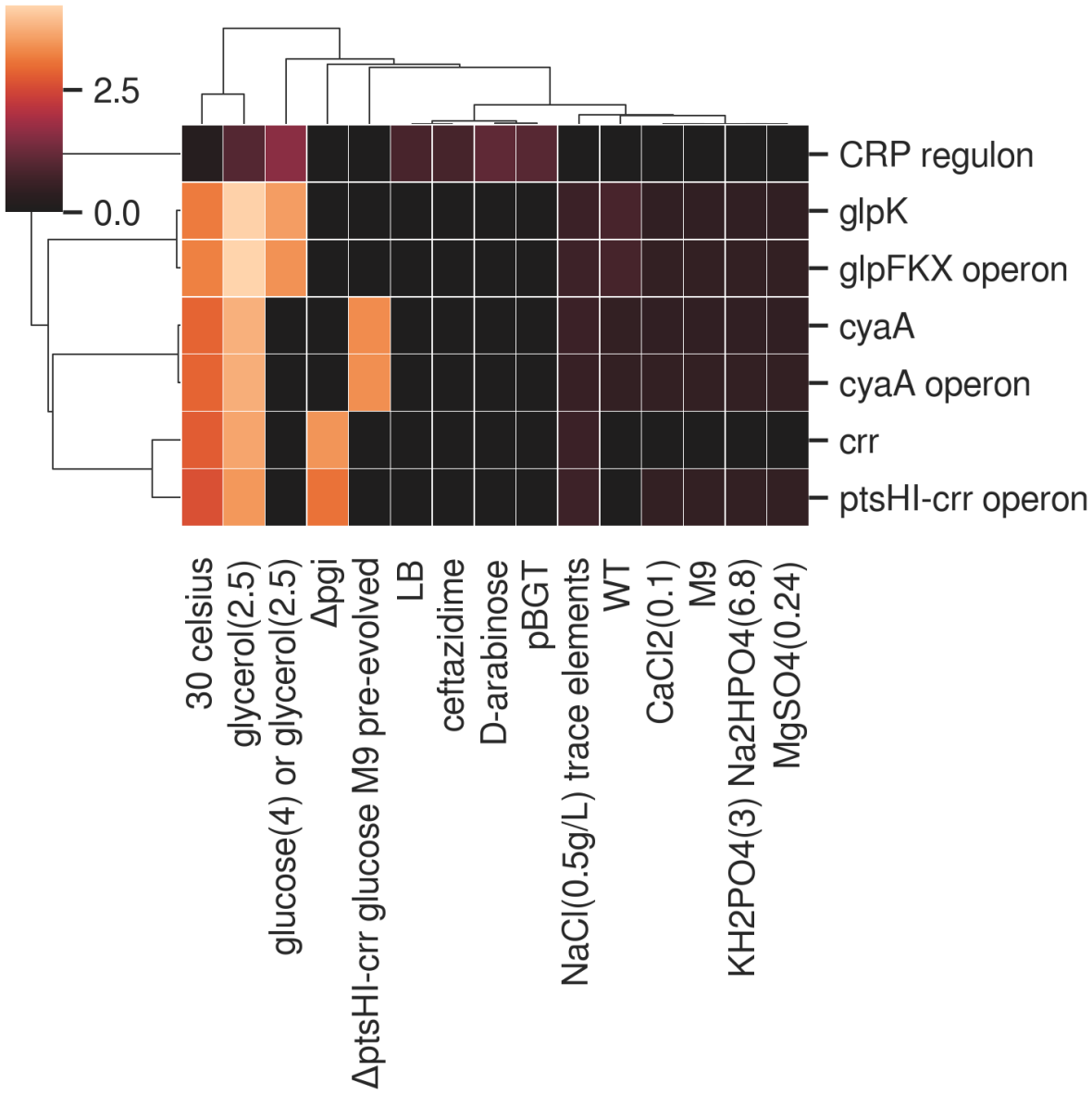


Figure C.16: A clustermap of the odds ratios for features of interest significantly associated with glycerol as a carbon source.

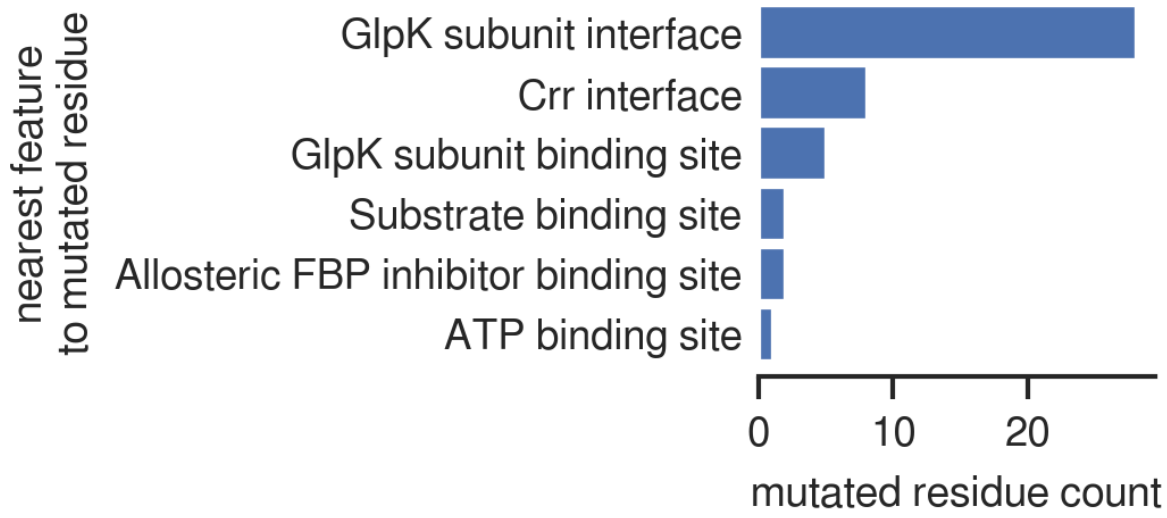


Figure C.17: The amount of unique mutated residues nearest to those residues describing functional domains on GlpK.

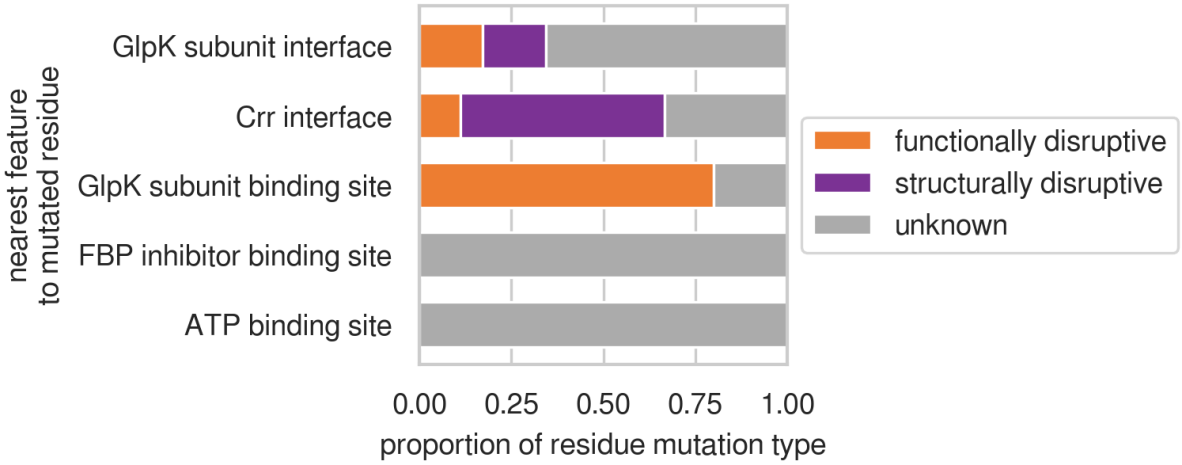
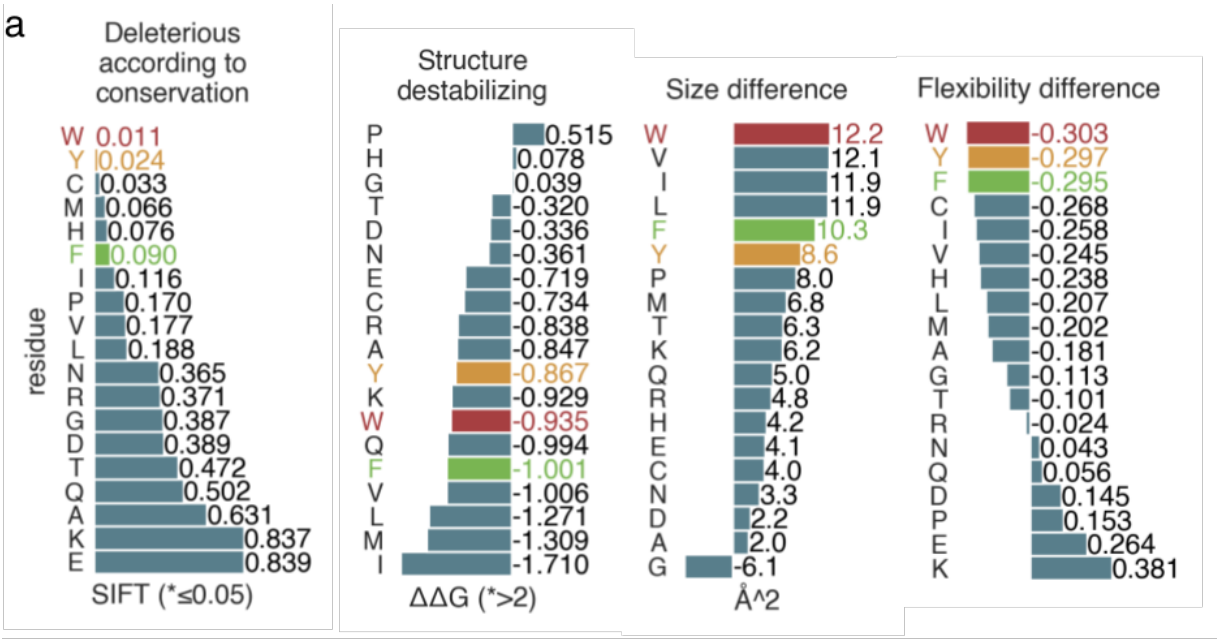


Figure C.18: The proportion of predicted mutation effects for all mutations to residues nearest to those residues describing functional domains on GlpK.



b

residue	SIFT (*≤0.05)	ΔΔG (*>2)	WT size difference	WT flexibility difference	WT hydrophobicity difference	size type	chain type	polar	charged	basic	acidic	common in turn	common in helix	common in sheet
S	1.000	0.000	0	0	0	tiny		yes				yes		
W	0.011	-0.935	12.2	-0.303	-0.1	bulky	aromatic						yes	
Y	0.024	-0.867	8.56	-0.297	-0.5	bulky	aromatic						yes	
F	0.090	-1.001	10.33	-0.295	3.6	bulky	aromatic						yes	
I	0.116	-1.710	11.93	-0.258	5.3	bulky	aliphatic						yes	yes
V	0.177	-1.006	12.1	-0.245	5	bulky	aliphatic						yes	
L	0.188	-1.271	11.93	-0.207	4.6	bulky	aliphatic						yes	
P	0.170	0.515	7.96	0.153	-0.8	small						yes		
N	0.365	-0.361	3.35	0.043	-2.7	small		yes						
D	0.389	-0.336	2.21	0.145	-2.7	small		yes	yes		yes	yes		
C	0.033	-0.734	3.99	-0.268	3.3	tiny								
G	0.387	0.039	-6.07	-0.113	0.4	tiny						yes		
T	0.472	-0.320	6.3	-0.101	0.1	tiny		yes						
A	0.631	-0.847	2.03	-0.181	2.6	tiny	aliphatic						yes	
M	0.066	-1.309	6.78	-0.202	2.7									
H	0.076	0.078	4.22	-0.238	-2.4		aromatic	yes	yes	yes				
R	0.371	-0.838	4.81	-0.024	-3.7			yes	yes	yes				
Q	0.502	-0.994	4.98	0.056	-2.7			yes					yes	
K	0.837	-0.929	6.24	0.381	-3.1			yes	yes	yes				
E	0.839	-0.719	4.1	0.264	-2.7			yes	yes		yes		yes	

Figure C.19: (a) The ranking of residues substitutions according to residue properties for GlpK's residue represented by amino acid 59.(b) A description of multiple residue substitution properties for GlpK's residue represented by amino acid 59.

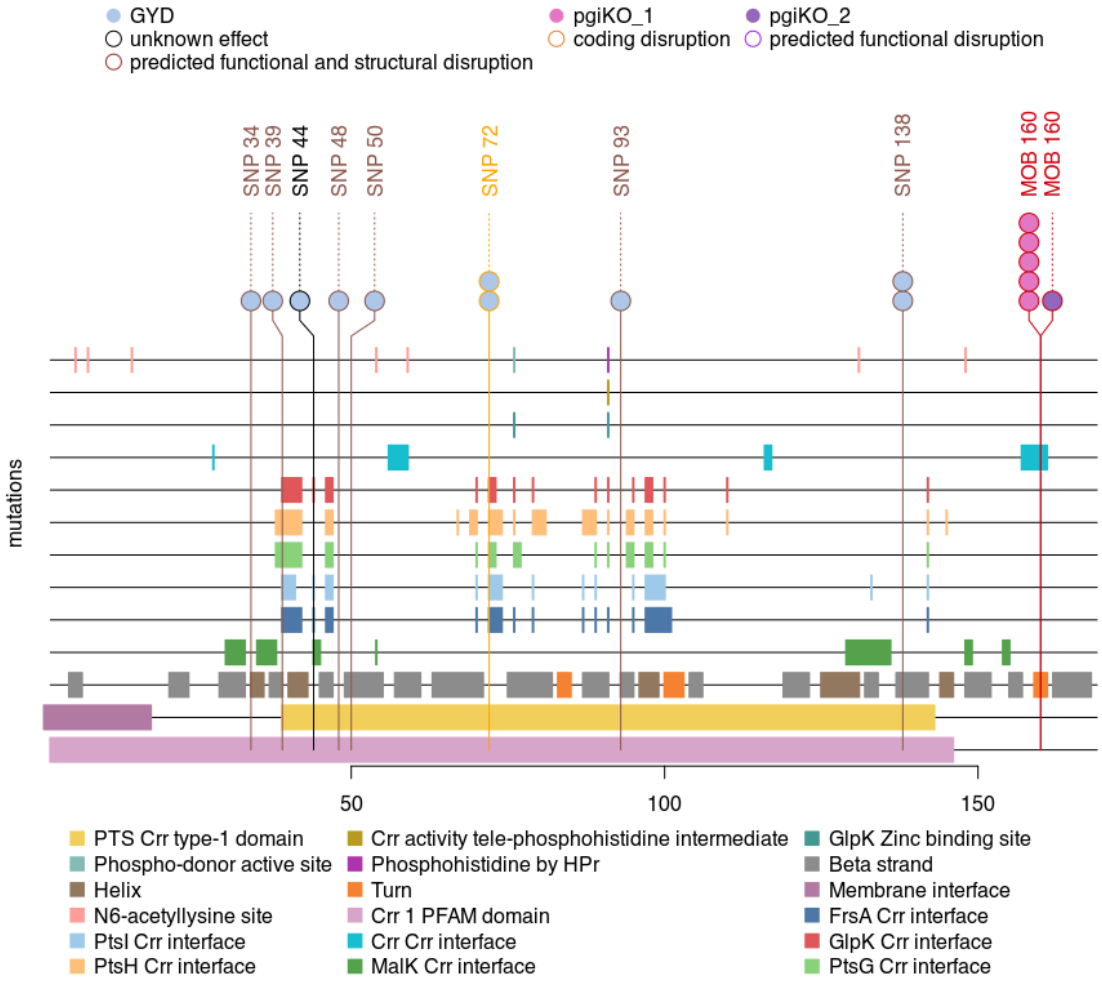


Figure C.20: A mutation needle plot for mutated amino acids across Crr's amino acid chain.

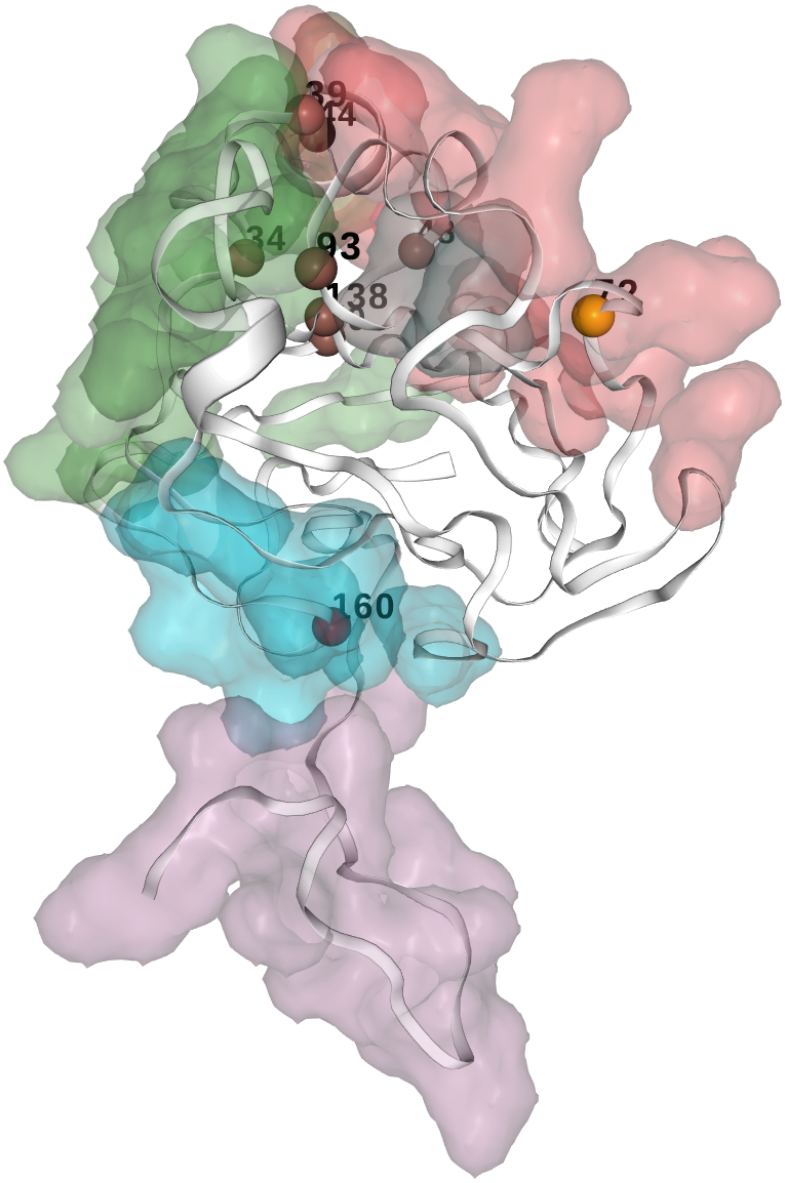


Figure C.21: Crr's 3D structure and mutated residues. The residue chain and transparent surfaces are colored according to the legend of the corresponding mutation needle plot. Mutations are represented by a small opaque sphere with a value representing their amino acid position on the corresponding mutation needle plot. The color of the mutation's sphere corresponds to the mutation's predicted effect as described by the legend on the corresponding mutation need plot. The transparent sphere centered on the mutations opaque sphere represents the number of mutations with a specific predicted effect on that position. The angle shown illustrates how the mutations cluster in 3D space and how the different surfaces of the 3D structure are used as different interfaces.

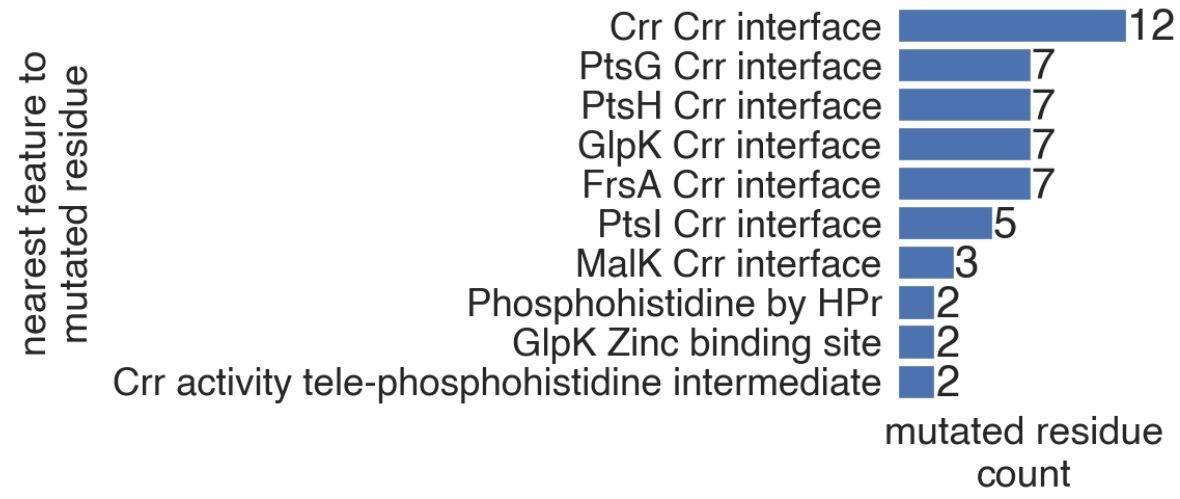


Figure C.22: The amount of unique mutated residues nearest to those residues describing functional domains on Crr.

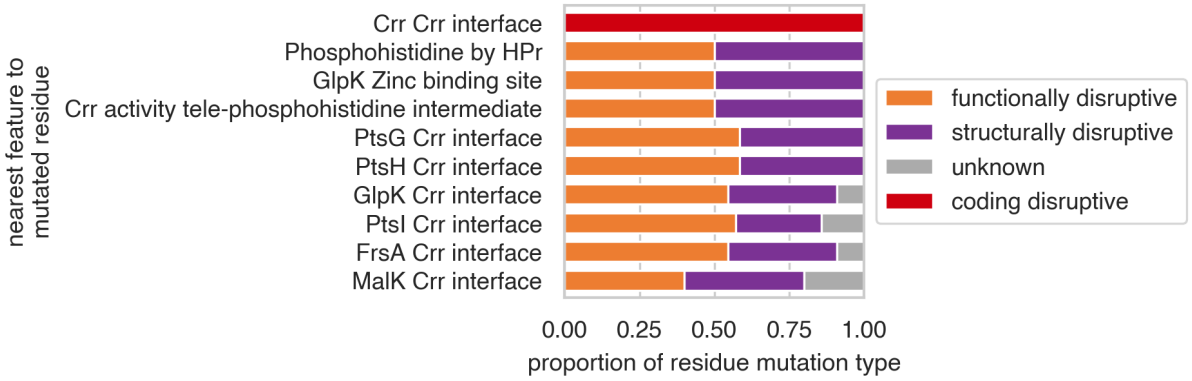


Figure C.23: The proportion of predicted mutation effects for all mutations to residues nearest to those residues describing functional domains on Crr.

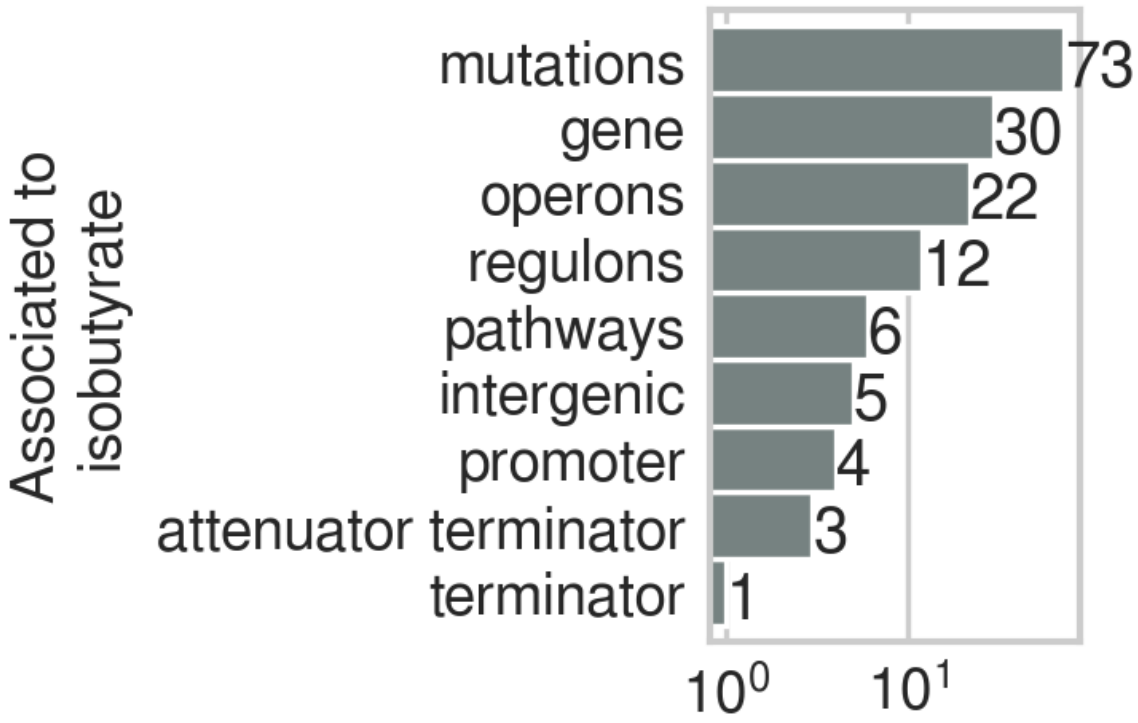


Figure C.24: The number of mutations and unique mutated features significantly associated with the supplement of isobutyric acid.

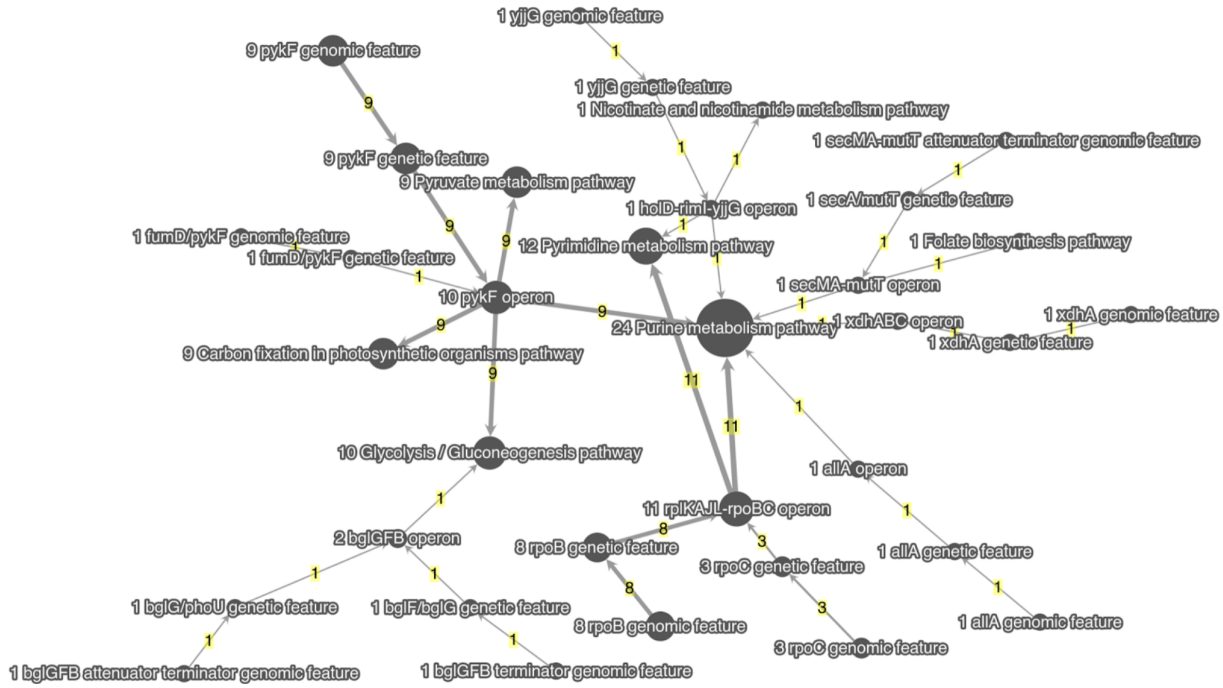


Figure C.25: A sub-network of mutated features associated with the isobutyric acid supplement and includes the Purine metabolism pathway demonstrates mutations to *pykF* and *rpoBC* being the primary contributors to mutation convergence on the Purine metabolism pathway. The nodes of the network represent features annotated on the genome and the edges represent connections between features. The weight and directionality of an edge represents the amount of smaller mutated features that can be represented by the larger feature. The weight of a node represents the amount of mutated features that converge on the specific genome annotation.

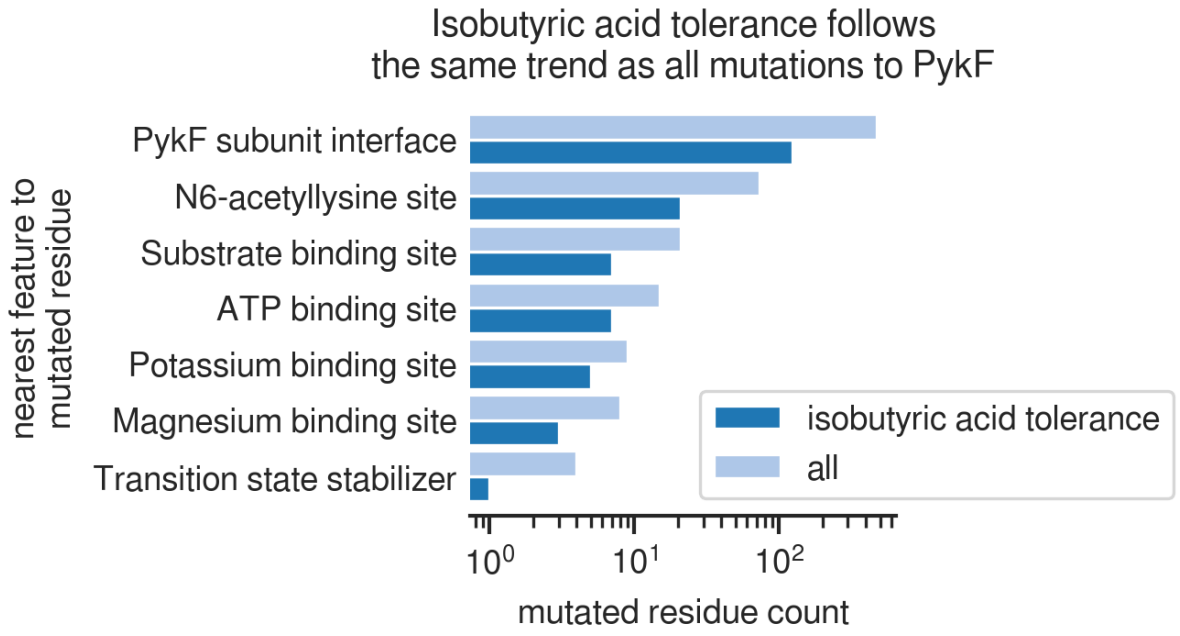


Figure C.26: The amount of unique mutated residues nearest to those residues describing functional domains on PykF. The mutated residues for all mutations and just those associated with isobutyric acid are shown.

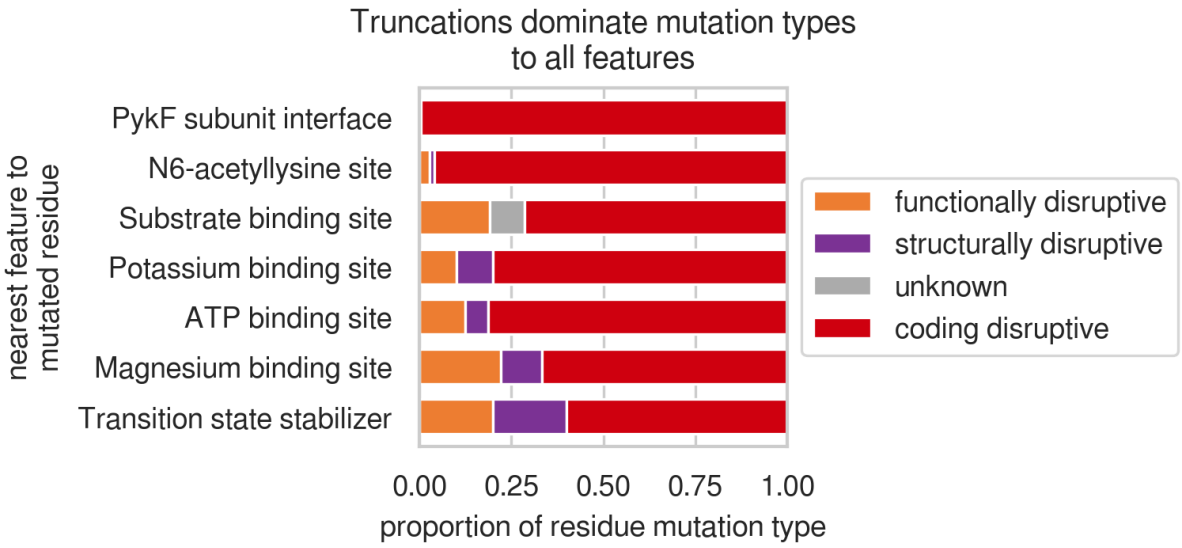


Figure C.27: The proportion of predicted mutation effects for all mutations to residues nearest to those residues describing functional domains on PykF.

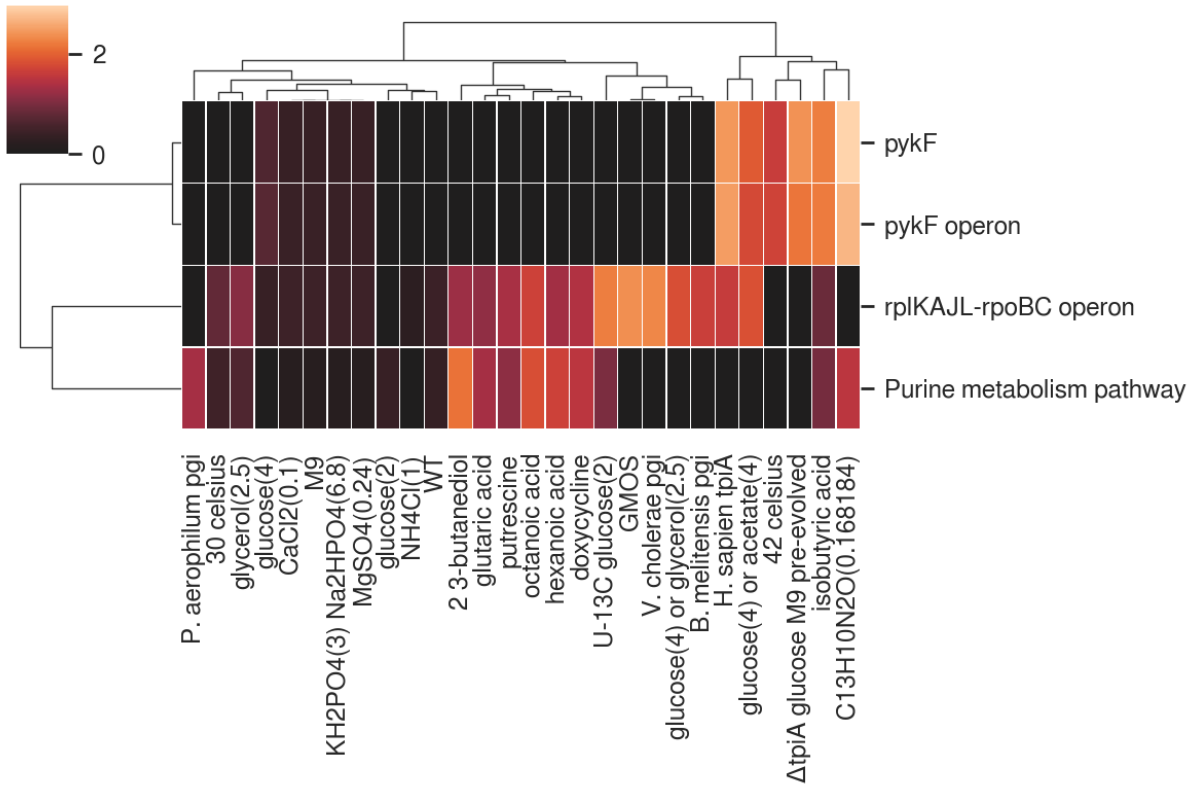


Figure C.28: A clustermap of the odds ratios for features of interest significantly associated with the supplement of isobutyric acid.

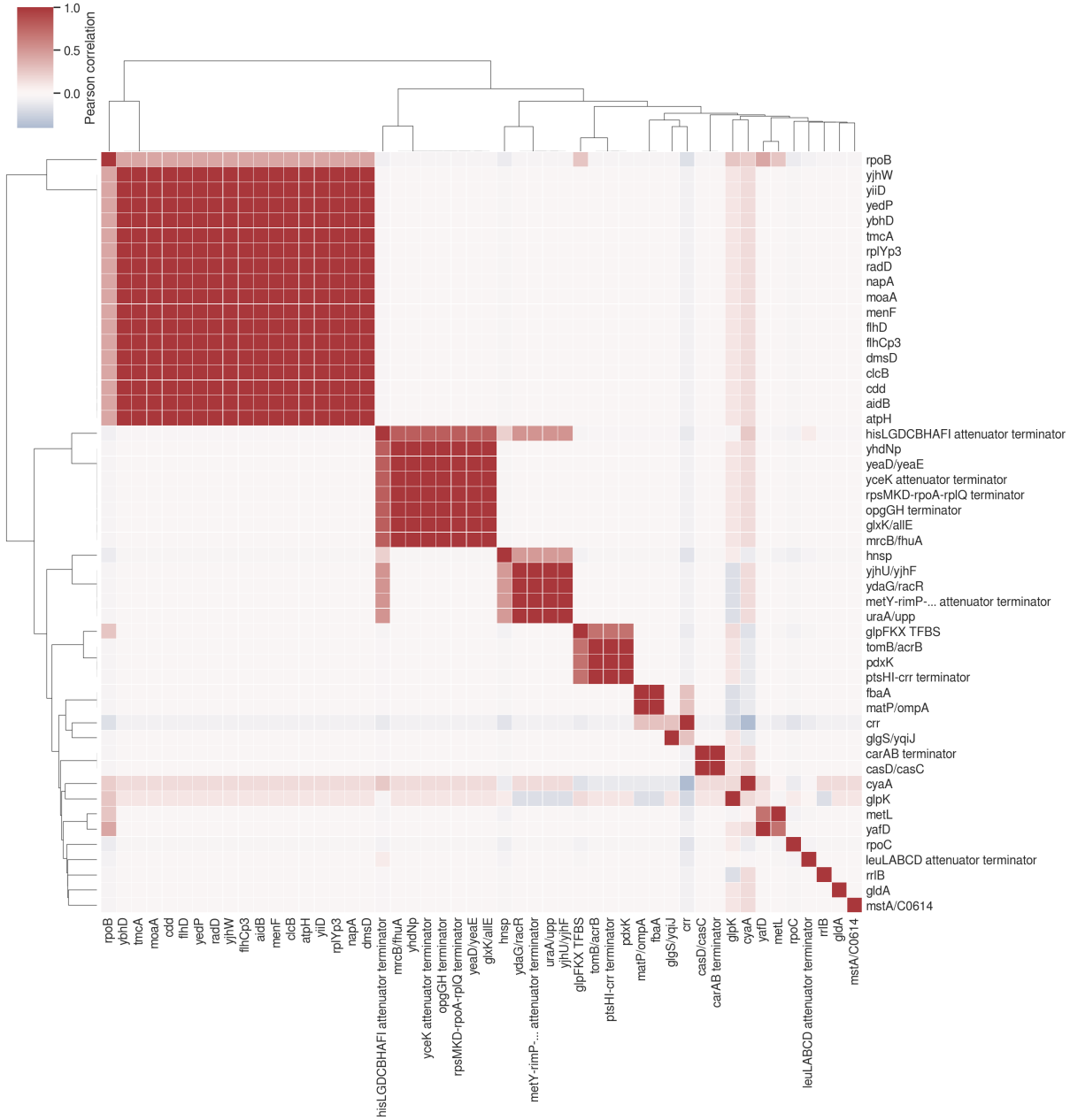


Figure C.29: A clustermap of the correlation between mutated genomic features significantly associated with glycerol as a carbon source.

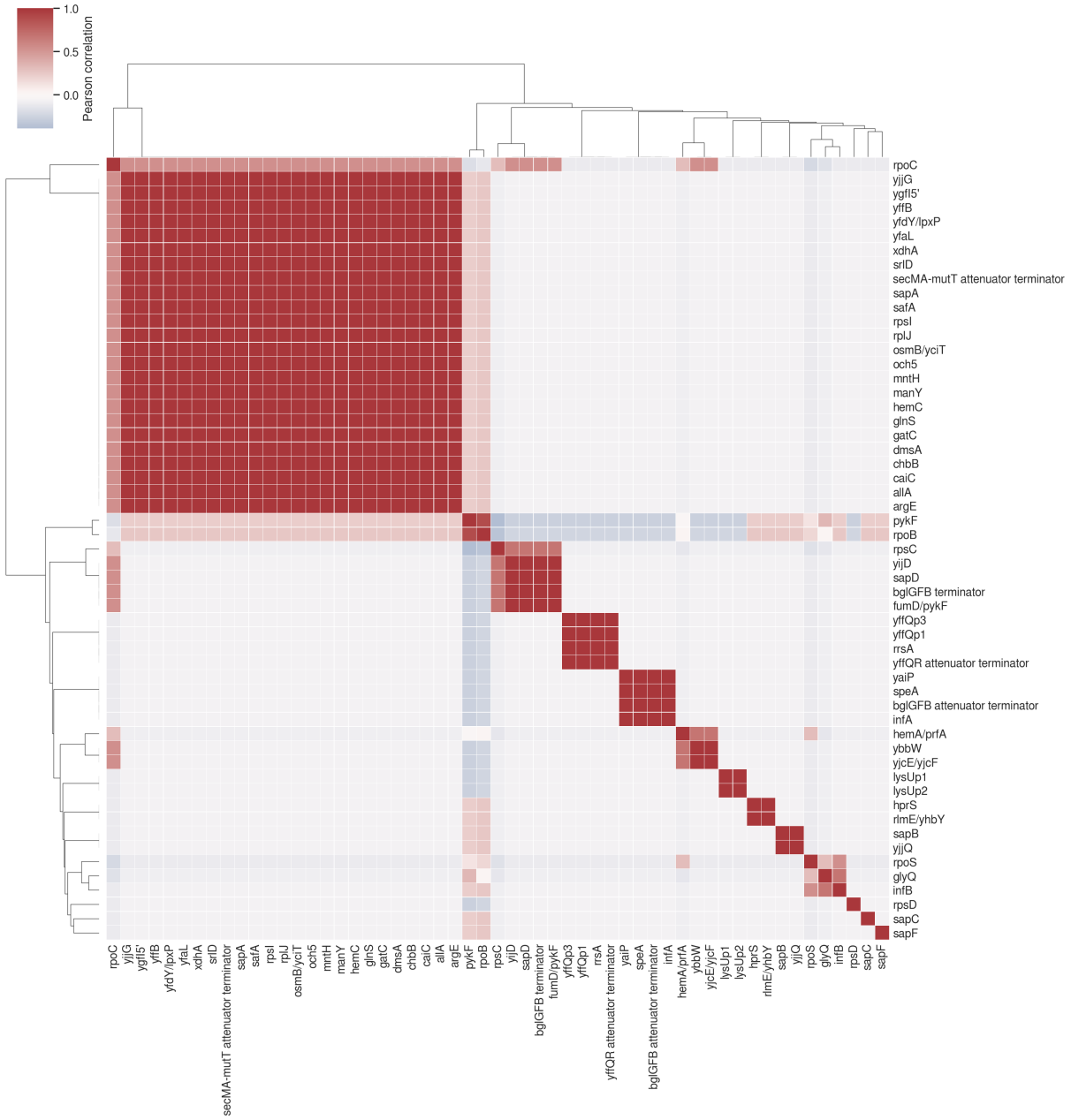


Figure C.30: A clustermap of the correlation between mutated genomic features significantly associated with isobutyric acid supplementation.

Table C.5: Table of mutated genomic features significantly associated with glycerol as a carbon source or linked to a larger feature with this association.

RegulonDB ID	genomic feature name	mutations
ECK120000391	glpK	30
ECK120000166	cyaA	21
ECK120000161	crr	10
ECK125144241	hisLGDCBHAFI attenuator terminator	6
ECK120000885	rpoB	5
ECK120000886	rpoC	4
ECK120010201	hnsp	4
ECK125143560	leuLABCD attenuator terminator	3
ECK125143900	yceK attenuator terminator	2
ECK120001548/ECK120002618	mstA/C0614	2
ECK120013464	glpFKX TFBS	2
ECK120010871	rpsMKD-rpoA-rplQ terminator	2
ECK120033617	yhdNp	2
ECK120010783	opgGH terminator	2
ECK125144631	metY-rimP-nusA-infB-rbfA-truB-rpsO-pnp attenuator terminator	2

Table C.5: Table of mutated genomic features significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120000838	rhsB	2
ECK120001091	fimZ	2
ECK120000583	metL	2
ECK120002891/ECK120002892	glxK/allE	2
ECK120000598/ECK120000296	mrcB/fhuA	1
ECK120001594	yafD	1
ECK120001029/ECK120030090	tyrB/yjbS	1
ECK120033723	yfdXp	1
ECK120001833	gldA	1
ECK120000276	fbaA	1
ECK120002835	yaiT	1
ECK120002403/ECK120000662	matP/ompA	1
ECK120002341	gntK	1
ECK120003553/ECK120003554	yeaD/yeaE	1
ECK120002520	rrlB	1
ECK125136882	insHp7	1
ECK125136879	mppAp5	1
ECK125136881	mppAp13	1
ECK125136880	mppAp9	1

Table C.5: Table of mutated genomic features significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120034219	ynaIp	1
ECK125136878	mppAp4	1
ECK125143994	ynaI attenuator terminator	1
ECK125143996	insH-4 attenuator terminator	1
ECK120003260	mppA	1
ECK120003259	pgrR	1
ECK120003261	ynaI	1
ECK120004249	gspE	1
ECK125136750	yceQp6	1
ECK120001349/ECK120004144	glgS/yqiJ	1
ECK120003274/ECK120003275	ydaG/racR	1
ECK120000903	rpsM	1
ECK120011926	hns TFBS	1
ECK120001044/ECK120003456	uidA/uidR	1
ECK120002998	ybhD	1
ECK120001543	moaA	1
ECK120003438	dmsD	1
ECK120003439	clcB	1
ECK125137156	flhCp3	1

Table C.5: Table of mutated genomic features significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120000314	flhD	1
ECK120003634	yedP	1
ECK120000133	cdd	1
ECK125137284	rplYp3	1
ECK120001961	radD	1
ECK120001976	napA	1
ECK120002230	menF	1
ECK120003870	tmcA	1
ECK120000101	atpH	1
ECK120001784	yiiD	1
ECK120001745	aidB	1
ECK120016925	yjhW	1
ECK120011932	hns TFBS	1
ECK120002346/ECK120002347	mdaB/ygiN	1
ECK120002200	yadK	1
ECK120002285/ECK120001647	tomB/acrB	1
ECK120010797	ptsHI-crr terminator	1
ECK120003833	pdxK	1
ECK120000160	crp	1

Table C.5: Table of mutated genomic features significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK125257016	carAB terminator	1
ECK120003992/ECK120003993	casD/casC	1
ECK120002031/ECK120001303	uraA/upp	1
ECK120004404/ECK120002319	yjhU/yjhF	1
ECK120002259	dgcE	1

Table C.6: Table of mutated operons significantly associated with glycerol as a carbon source or linked to a larger feature with this association.

operon name	mutations
glpFKX	32
cyaA	21
ptsHI-crr	11
rplKAJL-rpoBC	9
hisLGDCBHAFI	6
hns	6
leuLABCD	3
rpsMKD-rpoA-rplQ	3
ynaI	3
metY-rimP-nusA-infB-rbfA-truB-rpsO-pnp	2
fimZ	2

Table C.6: Table of mutated operons significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

operon name	mutations
yceK	2
opgGH	2
yhdN-zntR	2
rhsB	2
metBL	2
gntRKU	1
rrsB-gltT-rrlB-rrfB	1
mppA	1
pgrR	1
fhuACDB	1
tyrB	1
yfdX-frc-oxc-yfdVE	1
ptsA-fsaB-gldA	1
epd-pgk-fbaA	1
yaiT	1
insH-4	1
glgS	1
gspCDEFGHIJKLMO	1
upp-uraA	1

Table C.6: Table of mutated operons significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

operon name	mutations
yafDE	1
casABCDE12	1
napFDAGHBC-ccmABCDEFGH	1
uidABC	1
ybhD	1
moaABCDE	1
ynfEFGH-dmsD	1
clcB	1
flhDC	1
yedP	1
cdd	1
radD	1
menFDHBCE	1
carAB	1
ypfJ-tmcA	1
atpIBEFHAGDC	1
yihXY-dtd-yiiD	1
aidB	1
insO-2-yjhWV	1

Table C.6: Table of mutated operons significantly associated with glycerol as a carbon source or linked to a larger feature with this association, continued.

operon name	mutations
yadMLKC	1
tomB-hha	1
pdxK	1
crp	1
dgcE	1

Table C.7: Table of mutated regulons significantly associated with glycerol as a carbon source.

regulon name	mutations
CRP	73
GlpR	32
Cra	12
Fis	11
Mlc	10
H-NS	10
NagC	10
GadX	6
CspA	6
LeuO	4
FNR	4

Table C.7: Table of mutated regulons significantly associated with glycerol as a carbon source, continued.

regulon name	mutations
Fur	3
ArgR	3
IHF	2
PhoP	2
ModE	2
Lrp	2
NarL	2
SlyA	2
HypT	2
MetJ	2
TyrR	1
Ada	1
PdhR	1
BasR	1
CpxR	1
PepA	1
GntR	1
RutR	1
PurR	1

Table C.7: Table of mutated regulons significantly associated with glycerol as a carbon source, continued.

regulon name	mutations
FlhDC	1
EvgA	1
AidB	1
FliZ	1
NarP	1
IscR	1
CytR	1
NsrR	1
HdfR	1
LrhA	1
QseB	1
YjjQ	1
MatA	1
AcrR	1
RcsAB	1
OmpR	1
CueR	1
UxuR	1
UidR	1

Table C.7: Table of mutated regulons significantly associated with glycerol as a carbon source, continued.

regulon name	mutations
IdnR	1

Table C.8: Table of mutated regulons significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association.

RegulonDB ID	genomic feature name	mutations
ECK120000795	pykF	9
ECK120000885	rpoB	8
ECK120000503	rpoS	3
ECK120000886	rpoC	3
ECK120000893	rpsC	2
ECK120000402	glyQ	2
ECK120000420/ECK120000752	hemA/prfA	2
ECK120002019	yjjG	1
ECK125144826	bglGFB attenuator terminator	1
ECK120000862	rplJ	1
ECK120000497	infA	1
ECK120002675	sapB	1
ECK120001259	argE	1
ECK120000948	speA	1
ECK120000894	rpsD	1

Table C.8: Table of mutated genomic features significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120000422	hemC	1
ECK120002829	yaiP	1
ECK120002888	ybbW	1
ECK120004440	yjjQ	1
ECK125135301	och5	1
ECK120000498	infB	1
ECK120001868/ECK120001869	yjcE/yjcF	1
ECK120002674	sapC	1
ECK120010524	lysUp2	1
ECK120010265	lysUp1	1
ECK125137387	yffQp1	1
ECK125137388	yffQp3	1
ECK125144381	yffQR attenuator terminator	1
ECK120000899	rpsI	1
ECK120004048	xdhA	1
ECK125272704	ygfI5'	1
ECK120000672/ECK120003233	osmB/yciT	1
ECK120001463/ECK120002385	rlmE/yhbY	1
ECK120002183	sapD	1

Table C.8: Table of mutated genomic features significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120003492/ECK120000795	fumD/pykF	1
ECK120010798	bglGFB terminator	1
ECK120001362	yijD	1
ECK120001511	caiC	1
ECK125143570	secMA-mutT attenuator terminator	1
ECK120002883	allA	1
ECK120000383	glnS	1
ECK120000226	dmsA	1
ECK120002184	sapF	1
ECK120003642	hprS	1
ECK120002676	sapA	1
ECK120003379	safA	1
ECK120000136	chbB	1
ECK120000561	manY	1
ECK120002276	gatC	1
ECK120002399	yfaL	1
ECK120003815/ECK120003816	yfdY/lpxP	1
ECK120003828	mntH	1

Table C.8: Table of mutated genomic features significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association, continued.

RegulonDB ID	genomic feature name	mutations
ECK120001131	yffB	1
ECK120000960	srlD	1
ECK120002526	rrsA	1

Table C.9: Table of mutated operons significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association.

operon name	mutations
rplKAJL-rpoBC	12
pykF	10
sapABCDF	5
nlpD-rpoS	3
lysU	2
rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ	2
bglGFB	2
glyQS	2
hemA-prfA-prmC	2
infA-serW	1
rplM-rpsI	1
hemCDXY	1

Table C.9: Table of mutated operons significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association, continued.

operon name	mutations
argE	1
holD-rimI-yjjG	1
yaiP	1
yjjQ-bglJ	1
speAB	1
rpsMKD-rpoA-rplQ	1
yffQR	1
srlAEBD-gutM-srlR-gutQ	1
gcl-hyi-glxR-ybbW-allB-ybbY-glxK	1
metY-rimP-nusA-infB-rbfA-truB-rpsO-pnp	1
xdhABC	1
yfdY	1
yffB-dapE-ypfN	1
dmsABC	1
yhbY	1
rlmE-ftsH	1
fabR-yijD	1
caiTABCDE	1
secMA-mutT	1

Table C.9: Table of mutated operons significantly associated with toxic isobutyrate supplementation or linked to a larger feature with this association, continued.

operon name	mutations
allA	1
glnS	1
osmB	1
mntH	1
safA-ydeO	1
chbBCARFG	1
manXYZ	1
gatYZABCD	1
yfaL	1
lpxP	1
hprRS	1
rrsA-ileT-alaT-rrlA-rrfA	1

Table C.10: Table of mutated pathways significantly associated with toxic isobutyrate supplementation.

pathway name	mutations
Purine metabolism	24
Pyrimidine metabolism	12
Glycolysis / Gluconeogenesis	10
Pyruvate metabolism	9

Table C.10: Table of mutated pathways significantly associated with toxic isobutyrate supplementation, continued

pathway name	mutations
Carbon fixation in photosynthetic organisms	9
Aminoacyl-tRNA biosynthesis	5
Arginine and proline metabolism	2
Folate biosynthesis	1
Fructose and mannose metabolism	1
Porphyrin and chlorophyll metabolism	1
Nicotinate and nicotinamide metabolism	1

Table C.11: Table of mutated regulons significantly associated with toxic isobutyrate supplementation.

regulon name	mutations
CRP	12
Cra	11
ArcA	9
FNR	8
Fis	6
H-NS	5
Fur	4
GadX	3
LeuO	3

Table C.11: Table of mutated regulons significantly associated with toxic isobutyrate supplementation, continued.

regulon name	mutations
MqsA	3
IHF	3
RcsB	2
ArgR	2
NagC	2
Lrp	2
AllR	2
RcsB-BglJ	2
StpA	2
OxyR	1
PurR	1
CaiF	1
SrlR	1
GutM	1
YdeO	1
NarL	1
MntR	1
GatR	1
Mlc	1

Table C.11: Table of mutated regulons significantly associated with toxic isobutyrate supplementation, continued.

regulon name	mutations
ModE	1
ChbR	1
PhoP	1
EvgA	1

C.2 References

1. Zhao, D., Yuan, S., Xiong, B., Sun, H., Ye, L., Li, J., Zhang, X. & Bi, C. Development of a fast and easy method for Escherichia coli genome editing with CRISPR/Cas9. en. *Microb. Cell Fact.* **15,** 205. ISSN: 1475-2859 (Dec. 2016).

2. Mehrer, C. R., Incha, M. R., Politz, M. C. & Pfleger, B. F. Anaerobic production of medium-chain fatty alcohols via a β-reduction pathway. en. *Metab. Eng.* **48,** 63–71. ISSN: 1096-7176, 1096-7184 (July 2018).

3. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. ssbio: a Python framework for structural systems biology. en. *Bioinformatics* **34,** 2155–2157. ISSN: 1367-4803, 1367-4811 (June 2018).

4. Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A. & Beltrao, P. A resource of variant effect predictions of single nucleotide variants in model organisms. en. *Mol. Syst. Biol.* **14,** e8430. ISSN: 1744-4292 (Dec. 2018).

5. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. en. *Nucleic Acids Res.* **47,** D506–D515. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

6. Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I. & Karp, P. D. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. en. *Nucleic Acids Res.* **45,** D543–D550. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

7. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. The Pfam protein families database in 2019. en. *Nucleic Acids Res.* **47,** D427–D432. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

8. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908. ISSN: 1087-0156, 1546-1696 (Oct. 2017).

9. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* **47,** D1164–D1171. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

10. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9,** 90–95. ISSN: 1521-9615 (May 2007).

11. Waskom, M., Gelbart, M., Botvinnik, O., Ostblom, J., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Warmenhoven, J., Cole, J. B., de Ruiter, J., Vanderplas, J., Hoyer, S., Pye, C., Miles, A., Swain, C., Meyer, K., Martin, M., Bachant, P., Quintero, E., Kunter, G., Villalba, S., Brian, Fitzgerald, C., Evans, C., Williams, M. L.,

O'Kane, D., Yarkoni, T. & Brunner, T. *mwaskom/seaborn: v0.11.1 (December 2020)* Dec. 2020.

12. Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O. & Bader, G. D. Cytoscape.js: a graph theory library for visualisation and analysis. en. *Bioinformatics* **32,** 309–311. ISSN: 1367-4803, 1367-4811 (Jan. 2016).

13. Ou, J. & Zhu, L. J. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. en. *Nat. Methods* **16,** 453–454. ISSN: 1548-7091, 1548-7105 (June 2019).

14. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. en. *Bioinformatics* **32,** 2847–2849. ISSN: 1367-4803, 1367-4811 (Sept. 2016).

15. Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A. & Rose, P. W. NGL viewer: web-based molecular graphics for large complexes. en. *Bioinformatics* **34,** 3755–3758. ISSN: 1367-4803, 1367-4811 (Nov. 2018).