Title

Unraveling the sensory systems of cells and regulation of gene expression: characterization, dissemination, &amp; evolution of iModulons

Author

Rychel, Kevin William

Publication Date

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Unraveling the sensory systems of cells and regulation of gene expression: characterization, dissemination, and evolution of iModulons

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

Doctor of Philosophy

in

Bioengineering

by

Kevin William Rychel

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jeff Hasty
Professor Joseph Pogliano
Professor Karsten Zengler

2023

This dissertation of Kevin William Rychel is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

For all those I’ve lost this past year:

To my best friend Jeff, thank you for the years of creativity, curiosity, and fun. I hope I can keep asking big questions and living life to the fullest.

To my grandma Frances, thank you for helping raise me with your generous heart. I hope I can keep making you proud.

To my friend Stephen, thank you for the excitement you brought to life and to science. I hope I can hold onto your feelings of awe and the beauty of discovery.

To my father-in-law Nate, thank you for sharing your joy and loved ones with me. I hope I can follow your example of how to love family and strangers alike.

See page *xiii*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I thank Dr. Bernhard Palsson. The first time that I heard him lecture, I was filled with an overwhelming sense of inspiration — his ability to articulate the importance of our current moment in science, his grand vision for the future, and his confidence in our group's ability to make a powerful impact were jaw-dropping. The lab he has built is truly exceptional; I couldn't have asked for a better place to learn, grow, and help to lead. One of Dr. Palsson's greatest strengths is his ability to make a student feel accomplished and appreciated, which really encouraged me to go the extra mile. I am lucky that he saw teaching and leadership abilities in me, and trusted me to develop those skills in important roles. I have learned so much from him, and I'm beyond proud to be a part of his legacy.

The entirety of this dissertation was made possible by advances made by my mentor and friend, Anand Sastry. The original development and rapid expansion of iModulons that he orchestrated required so much creativity, intelligence, hard work, and patience. Having the opportunity to gain such close and helpful mentorship from him and use his expertly built software was absolutely invaluable to me. His leadership is a model that I tried to emulate later in my PhD career, and I hope that I was able to impart a fraction of the impact that he had on me onto the students that came after me.

I have also received valuable help and mentorship from Dan Zielinski, who shaped my approach to data science through the courses he lectures, and then became a key person to lean on for all kinds of help and advice.

In my second aim, I learned how to develop a website. This would have potentially taken much, much longer, if not for the help and guidance of Patrick Phaneuf. His experience and assistance pointed me in a very fruitful direction. Patrick was also a huge help with any technical difficulty anyone in the lab had, which saved tons of headaches. He has been a wise mentor to me on several aspects of my other projects as well, which I really appreciate.

My third aim centers around laboratory evolution, and I was lucky enough to have Adam Feist leading both of those projects, which began before I even joined the group. His vision, expertise, and guidance have been essential. In a similar vein, I would like to thank Justin Tan and Ke Chen, who generated the strains I analyzed as part of this aim. They both left behind really excellent work, which I was excited to build upon. They were extremely helpful to me despite having moved on from the lab, which is very kind and greatly appreciated.

Computational biology cannot exist without wet lab biology. Ying Hefner and Richard Szubin both deserve a huge thank-you for all of their hard work and patience as experimentalists, carrying out the experiments and generating the data at the center of everything I do. They are also both such passionate individuals and are a joy to work with.

I would also like to thank my fellow students and postdocs in the lab. To name a few: Arjun Patel, Cameron Lamoureux, Chris Dalldorf, Saugat Poudel, Sidd Chauhan, Katherine Decker, Hyungyu Lim, Annie Yuan, Akanksha Rajput, and Ye Gao. It has been a pleasure to work together and learn from each of them. We have a friendly and helpful culture in the lab, which makes it such a great and successful place.

I would like to thank the students I have recruited or helped mentor. Heera Bajpe has helped to remind me of the joy of doing science, and it has been so rewarding to watch her grow and to see the fruits of her hard work. Jayanth Krishnan is an extremely bright and hardworking student, and he really stepped up to the challenge of updating and maintaining our software, which is so important for the continued success of iModulons. I also appreciate Amy Lou, Joshua Burrows, Griffith Hughes, and Archana Balasubramanian for participating in my subgroup meetings — I learn a lot just by staying up to date and making suggestions about your projects.

One of my favorite aspects of my PhD experience has been the chance to collaborate with people around the world. I am thankful to Omkar Mohite, Laurence Yang, Bram Kerssemakers, Josefin Johnsen, Elsayed Mohamed, Emre Özdemir, Lei Yang, Tobias Alter, Shannara Kayleigh Taylor Parkins, Christoffer Rode, Yujiro Hirose, André Birgy, and the many others who provided me with guidance, insight, data, collaboration opportunities, or simply friendship across borders.

I would also like to thank my committee. Dr. Pogliano was kind enough to help me in my first project using his *Bacillus* expertise. Dr. Hasty and Dr. Zengler both mentored students I collaborated with

(Arianna Miano and Juan Tibocha-Bonilla). All three experiences were wonderful, and helped me to see iModulons from a new perspective. I really appreciate their time and feedback.

Before entering my PhD, I was inspired to continue research in large part by Daniel Clough. It is rare for an undergraduate researcher to receive such friendly and intellectually-stimulating mentorship from a grad student, so I was extremely lucky to have worked with Dan and been welcomed into the research community with open arms.

More broadly, I need to thank my family and friends. I could not have gotten anywhere in life without my parents. My mom was my initial inspiration to pursue science, and her loving support has been a constant in my life despite the hardships that she has had to face. Both of my parents have set me up for success in more ways than I can imagine, and I hope that I can continue to make them proud. To my siblings and friends, you made this period of my life so much fun. Thanks for the distractions, and a special thank-you to those that helped to keep me focused.

I would also like to thank my fiancé, Sean. He has been so incredibly patient through this graduate school process, and it would have been completely miserable without him. His support has been so generous and came in so many forms, from cheering me up after long days to literally giving me the shirt off his back. I could not be more grateful to have him in my life, and I am so excited for our future together.

Some very important people in my life passed away during the last year of this work. I would like to thank each of them for the experiences I was able to have with them, and I have dedicated this work to their memory. Jeff Donaldson (1996-2022) was my best friend during my high school years. I wish I could let him know how much I valued his creative intellect; how growing up with him set me up to ask the big questions I have been chasing in this dissertation. Frances Sprowls (1929-2023) was my grandmother, and one of the most loving and generous people I have ever met. She showed me that “where there is a will, there is a way,” which I definitely leaned on during grad school. Stephen Calderon (1993-2023) was a great friend to me for the past year, and I loved sharing both science and music with him. He was full of pride and childlike excitement about everything scientific, which was incredibly inspiring to me. Nate Penn (1951-2023) was my father-in-law, and he spent his years bringing so much joy to everyone — from close loved ones to strangers on the street. He built the family I am so excited to join, and showed Sean how to be the incredible man who supported me throughout my journey. I wish I could share more time with each of these four amazing people.

I also want to thank those who I have shared grief with, especially: Katie Bertram, Jena Mogensen, the Donaldson family, the entire Michalak family (especially my mom), Samuel Gilbert, and the entire Penn family (especially Sean and Andrea Penn). I learned a lot about myself by sharing these difficult times with you, and I know we will all continue to grow together.

Finally, I would like to thank the funding sources which supported this work. My primary funder was the Novo Nordisk foundation, under grant numbers NNF10CC1016517 and NNF20CC0035580.

Chapter 2, in part, is a reprint of the material as appears in: **Rychel K**, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome. *Nat Commun.* 2020 Dec 11; 11(1):6338. The dissertation author was the primary author.

Chapter 3, in part, is a reprint of the material as appears in: **Rychel K,** Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res.* 2021 Jan 8; 49(D1):D112-D120.

Chapter 4, in part, is currently under review for publication: **Rychel K**, Tan J, Patel A, Lamoureux C, Hefner Y, Szubin R, Johnsen J, Mohamed ETT, Phaneuf PV, Anand A, Olson CA, Park JH, Sastry AV, Yang L, Feist AM, Palsson BO. Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance. 2023. The dissertation author is the primary author.

Chapter 5, in part, is currently being prepared for submission for publication: **Rychel K,** Chen K, Patel A, Phaneuf PC, Feist AM, Palsson BO. High heat tolerance mechanisms from the transcriptomes of laboratory evolved strains. 2023. The dissertation author is the primary author.

VITA

2018 Bachelor of Science in Engineering in Biomedical Engineering, Minor in Computer Science, University of Michigan Ann Arbor

2023 Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Rychel K, Chen K, Patel A, Olson CA, Sandberg TE, Gao Y, Xu S, Hefner Y, Szubin R, Feist AM, Palsson BO. Laboratory evolution reveals transcriptional mechanisms underlying thermal adaptation of *Escherichia coli*. *In Preparation.* 2023.

Rychel K, Tan J, Patel A, Lamoureux C, Hefner Y, Szubin R, Johnsen J, Mohamed ETT, Phaneuf PV, Anand A, Olson CA, Park JH, Sastry AV, Yang L, Feist AM, Palsson BO. Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance. *Submitted.* 2023.

Tibocha-Bonilla JD, Zuñiga C, Lekbua A, Lloyd C, **Rychel K**, Short K, Zengler K. Predicting stress response and improved protein overproduction in *Bacillus subtilis*. *NPJ Syst Biol Appl.* 2022 Dec 27; 8(1):50.

Yuan Y, Seif Y, **Rychel K**, Yoo R, Chauhan S, Poudel S, Al-Bulushi T, Palsson BO, Sastry AV. Pan-Genome Analysis of Transcriptional Regulation in Six *Salmonella enterica* Serovar Typhimurium Strains Reveals Their Different Regulatory Structures. *mSystems.* 2022 Dec 20; 7(6):e0046722.

Rajput A, Tsunemoto H, Sastry AV, Szubin R, **Rychel K**, Chauhan SM, Pogliano J, Palsson BO. Advanced transcriptomic analysis reveals the role of efflux pumps and media composition in antibiotic responses of *Pseudomonas aeruginosa. Nucleic Acids Res.* 2022 Sep 23; 50(17):9675-9688.

Lim HG, **Rychel K**, Sastry AV, Bentley GJ, Mueller J, Schindel HS, Larsen PE, Laible PD, Guss AM, Niu W, Johnson CW, Beckham GT, Feist AM, Palsson BO. Machine-learning from *Pseudomonas putida* KT2440 transcriptomes reveals its transcriptional regulatory network. *Metab Eng*. 2022 Jul; 72:297-310.

Yoo R, **Rychel K**, Poudel S, Al-Bulushi T, Yuan Y, Chauhan S, Lamoureux C, Palsson BO, Sastry A. Machine Learning of All *Mycobacterium tuberculosis* H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere.* 2022 Apr 27; 7(2):e0003322.

Decker KT*, Gao Y*, **Rychel K***, Al Bulushi T, Chauhan SM, Kim D, Cho BK, Palsson BO. proChIPdb: a chromatin immunoprecipitation database for prokaryotic organisms. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D1077-D1084.

Chauhan SM, Poudel S, **Rychel K**, Lamoureux C, Yoo R, Al Bulushi T, Yuan Y, Palsson BO, Sastry AV. Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile *Sulfolobus acidocaldarius*. *Front Microbiol.* 2021 Oct 27; 12:753521.

Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D112-D120.

Rychel K, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome. *Nat Commun.* 2020 Dec 11; 11(1):6338.

ABSTRACT OF THE DISSERTATION

Unraveling the sensory systems of cells and regulation of gene expression: characterization, dissemination, and evolution of iModulons

by

Kevin William Rychel

Doctor of Philosophy in Bioengineering

University of California San Diego, 2023

Professor Bernhard Ø. Palsson, Chair

Organisms use a complex transcriptional regulatory network (TRN) to sense their environments and alter their phenotypes in response, which is integral to their adaptation and survival. Transcriptomic datasets, which measure the complete activity of the TRN as expression values for each gene, have accumulated online in large numbers, and represent a goldmine of biological information. However, the large number of variables and uncharacterized global TRN structure present a significant challenge for interpreting these datasets. A recently developed machine learning method, iModulon analysis, can determine the global structure of TRNs by identifying co-regulated signals in compendia of diverse transcriptomic data. The output of this method is a set of independently modulated gene sets (iModulons) for the given organism, which can be characterized to discover transcriptional regulation mechanisms and other biological insights. Characterized iModulons enable a low dimensional, knowledge-enriched representation of the transcriptome, dramatically simplifying the analysis of the sensory and regulatory systems of cells. Here, we establish and expand the usefulness of iModulons in three aims: (1) we characterize the TRN of *Bacillus subtilis*, a model soil and gut bacterium, obtaining novel insights on a variety of functions including sporulation; (2) we develop iModulonDB.org, and online knowledgebase of iModulons, and we populate it with over ten characterized organisms from across the phylogenetic tree of life to create a widely applicable scientific knowledgebase; and (3) we combine iModulon analysis with DNA sequencing and other technologies to understand the evolution of stress tolerance for both oxidative and temperature stress. This body of work stands upon decades of bottom-up characterization of transcriptional regulators and analyzes years' worth of accumulating datasets to elucidate a new perspective on cellular function from the top-down. It reveals that a global understanding of the transcriptome in low dimensions is both possible and useful, and provides an online resource so that the scientific community can continue to mine it for new insights. Stress resistance mechanisms presented in Aim 3 are relevant to the evolution of pathogens and common biomanufacturing challenges. Taken together, this work establishes iModulons as a powerful and accessible tool for the understanding of gene expression.

Chapter 1. Introduction

1.1 A Brief History of Systems Biology

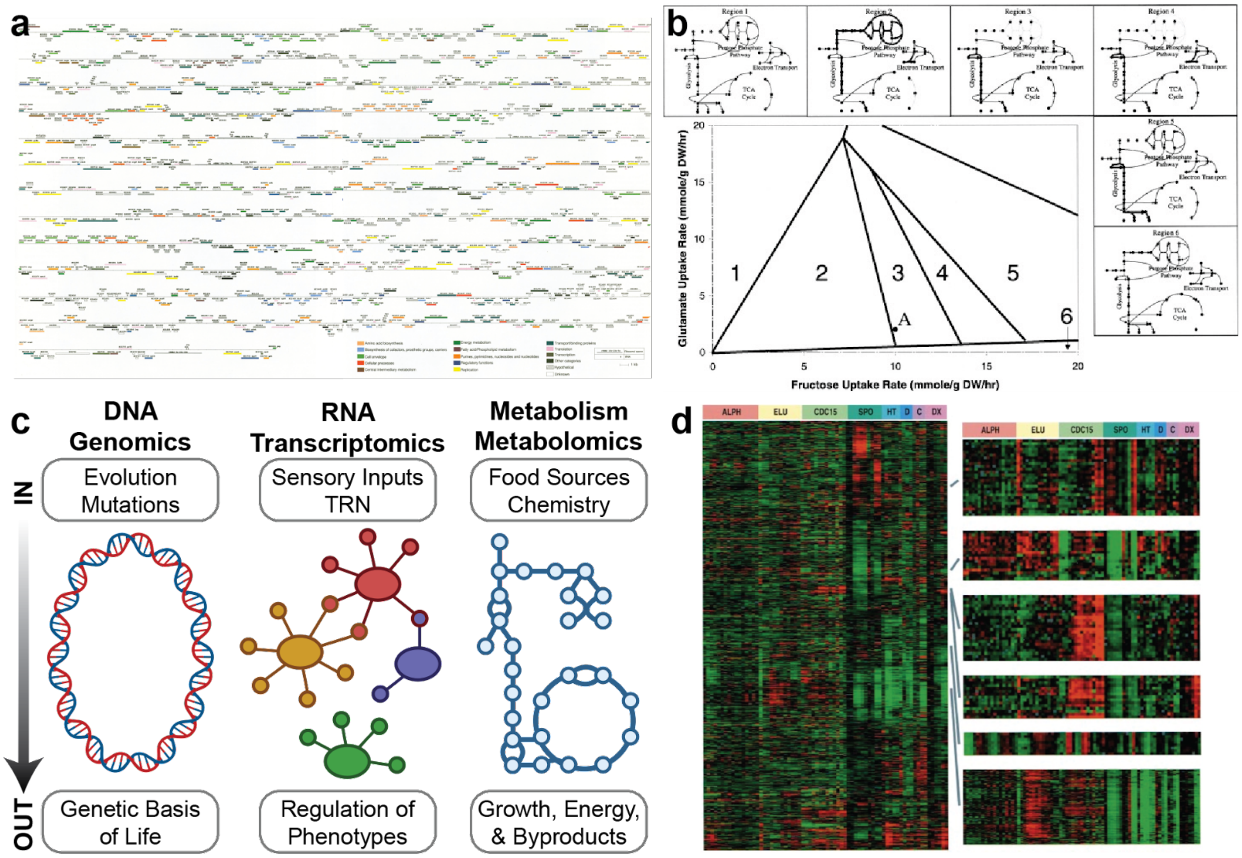


Figure 1.1: Development of global systems understanding in biology. **(a)** The first half of the first completely sequenced genome [1], with colored bars representing genes. All genes can be enumerated in full, heralding a new age of biological research. **(b)** Figure from the first genome-scale metabolic model [2], in which the many fluxes of the chemical networks along the top and right images can be easily interpreted by mapping to a lower-dimensional set of variables in the graph on the lower left. A major goal of systems biology is to interpret large datasets at the genome-scale, which is achieved here. **(c)** Depiction of the three relevant levels of systems biology, with the molecules, ‘omics types, inputs, and outputs shown. Genomics and metabolites are modeled in (a) and (b), and raw data for the transcriptomics is shown in (d), but revealing a more complete understanding of the transcriptional regulatory network (TRN) in the center is the goal of this dissertation. **(d)** Image of one of the first microarray transcriptomes [3], which is challenging to analyze due to the high number of variables and complexity of the underlying system.

When the first full genome was sequenced in 1995 [1] (**Figure 1.1a**), humans could, for the first time, begin to understand life at a "systems" level. Each part of the system of an organism (the genes) could be enumerated, and the way they function together could be understood from a new, global perspective. Understanding life in this way is of fundamental scientific interest, and the pursuit of genome-scale understanding has transformed the field of biology.

Edwards and Palsson built the first genome-scale metabolic model soon after the first genome, in 1999 [2] (**Figure 1.1b**). By arranging each of the known enzymatic reactions encoded in a genome into a chemical network reconstruction and applying flux balance analysis [4], global properties of the metabolic network could be revealed for the first time. Compared to its large number of reactions, the network itself took on a smaller number of states, which could be elucidated and interpreted. Genome-scale metabolic modeling has grown dramatically in the decades since this advance [5], and its applications have shown to be wide-reaching, including for the design of chemical production strains, identification of drug targets, and understanding of human disease [6], which motivates the continued development of genome-scale frameworks for interpreting biological systems.

Thus, two "levels" of biology were described from a global perspective. However, the understanding of biological systems is still incomplete. There exist additional levels between the genome and the metabolome, which have taken longer for the field to develop good models for **(Figure 1.1c)**. Each of the reactions in metabolic models are subject not only to changes in flux, as the early models describe, but also to changes in overall expression. The RNA and protein levels within cells change drastically based on sensory inputs, enabling the diverse function and adaptation of life we observe. To complete our global understanding of the cell, we must elucidate the transcriptional regulatory network (TRN) within it.

Also in the late 1990s, the first gene expression microarrays were developed. Called at the time a "new type of map," [3], they were able to quantify the expression of each gene at the genome-scale **(Figure 1.1d)**. We could now obtain a global view of RNA expression, one of the key middle layers of biological systems **(Figure 1.1c)**. Transcriptomic technology continued to improve, with RNA sequencing (RNAseq) arising in 2008 [7]. With these approaches, we could measure expression changes and quantify the complete output of the TRN. However, given the large number of gene variables and the striking complexity of the system, a good global model for analyzing this data remained elusive.

This dissertation presents a method for analyzing and interpreting transcriptomic data, which more broadly reveals a global, quantitative understanding of the sensory and regulatory systems of cells. It also seeks to connect results from this method across all three levels of biology described here. It advances systems biology by providing a widely applicable, scalable, and human interpretable perspective on the TRN, filling in the gaps in our understanding of biology.

1.2 Transcriptional Regulation & Data

Decades of research in transcriptional regulation have built up a large body of literature on its basic and specific functions. Upstream of each gene lies a promoter region, which may have binding sites for RNA polymerase and various transcription factors (TFs). TFs and other regulators activate or repress transcription of a gene based on cellular states [8], and together all regulators and their target genes comprise the TRN. By deleting or over-expressing TFs and measuring the changes in gene expression, and by identifying TF binding sites via chromatin immunoprecipitation, past research has defined regulons for each characterized TF [9]. A regulon is the set of all genes whose expression levels are expected to be controlled by the corresponding regulator based on a combination of experimental results. The field has taken a bottom-up approach, characterizing each regulator one at a time. Knowledge about each characterized regulon has been curated in useful databases such as BioCyc [10], RegulonDB [11], and *Subti*Wiki [12], which enable researchers to easily find information about their genes and regulators of interest.

Despite the wealth of knowledge, it remains difficult to analyze new transcriptomic data. The typical approach is to identify differentially expressed genes (DEGs) between conditions of interest [13], which requires researchers to parse hundreds or thousands of DEGs for any comparison. It is therefore difficult to interpret these results and nearly impossible to gain a global understanding from this method. A framework is needed which captures the structure of the underlying TRN and groups genes by their co-regulation. This would enable much more clarity in interpretation and quickly reveal a goldmine of information hidden in transcriptomic data.

Efforts to model the TRN have had some success, but significant room for improvement remains. Utilization of existing regulator annotations can be used to define a quantitative TRN, but these structures are often biased against uncharacterized genes and fail to accurately capture and predict transcriptomic data [14], [15]. Clustering and statistical inference methods do not share those limitations, but more effort is needed to establish a particular method which matches biological reality, robustly scales to large datasets, and is accessible to a wide range of researchers [16], [17]. A method that fits this description is put forward in this dissertation.

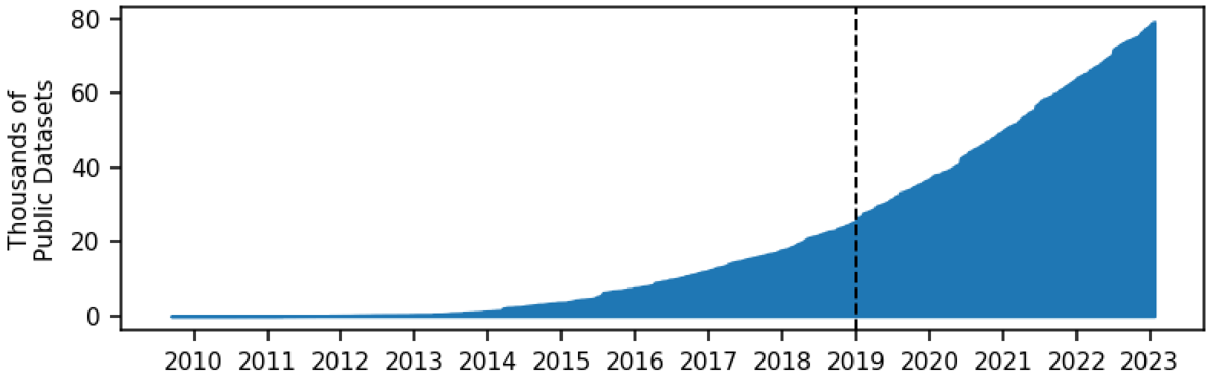


Figure 1.2: A “tsunami” of publicly available bacterial RNAseq profiles is available from the Sequence Read Archive (SRA) [18]. This graph shows the number of accumulated results from searching SRA for all public bacterial RNAseq data. The dotted vertical line represents the start date of this PhD dissertation.

While computational biologists have worked towards interpretation and modeling of transcriptomic datasets, the cost of their generation has dropped significantly and they have begun to accumulate in large numbers. The Sequence Read Archive (SRA) has experienced “explosive growth… a tsunami” of new publicly available transcriptomic data in the past decade **(Figure 1.2)** [18]. The large amount of available data presents an unprecedented opportunity: we can decipher the TRN from the top-down, by looking for patterns across all existing data for a strain. Since the bottom-up methods (characterizing regulators one at a time) have not been able to quantitatively capture the global structure at this level of biology, this new perspective is very promising. It would also be of wide interest to the many researchers who have contributed data to online databases, as new perspectives and insights can be revealed by re-analyzing their data at this scale.

To summarize, problems facing the transcriptomic analysis scientific community include: (i) the intractability of interpreting large numbers of DEGs, (ii) a lack of a quantitative structure that fits with both our existing knowledge of the TRN and real transcriptomic data, and (iii) a tsunami of available data has not been analyzed in a consistent way to reveal its top-down structure.

1.3 iModulons address transcriptomic analysis changes

In 2019, Sastry et al. published the first iModulon structure [19], which addresses the challenges described above. iModulons are *i*ndependently *modul*ated sets of genes which are computed from large transcriptomic compendia using the machine learning algorithm independent component analysis (ICA) [20]. ICA is a blind source separation algorithm originally developed to identify the independent signals underlying radar or mixed sound inputs, but can be applied to transcriptomic data to quantify underlying regulatory signals. The transcriptomic compendium used in the first iModulon paper is called PRECISE-278 (Precision RNAseq Compendium for Independent Signal Extraction with 278 samples). In this section, we will describe the general mathematical framework, some intuition on the underlying machine learning, and briefly summarize the early successes of iModulons.

1.3.1 Introduction to iModulons

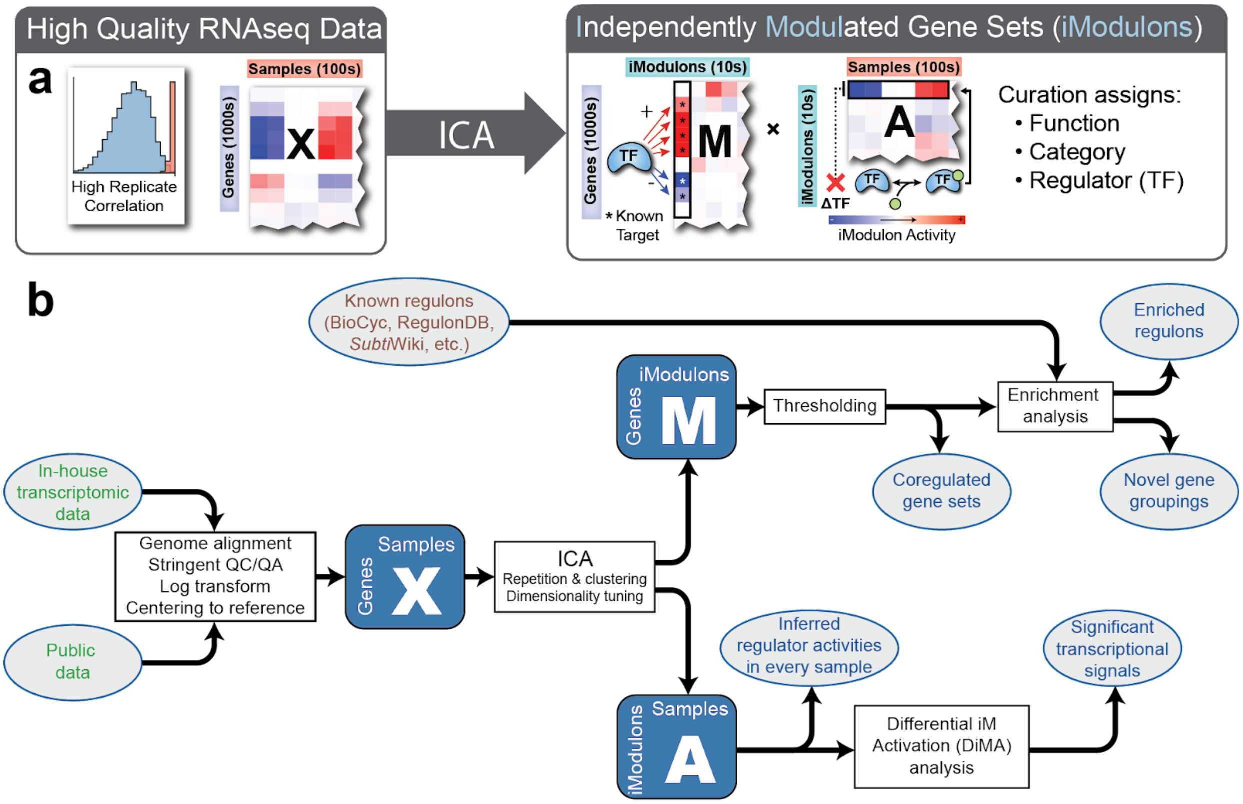


Figure 1.3: The iModulon analysis workflow identifies structure in transcriptomic datasets, which facilitates interpretation. **(a)** A simple representation of the iModulon workflow adapted from [21], which highlights the ICA equation **X** ~ **M** * **A**. The columns in **M** are iModulon gene weights, analogous to TF binding strengths, and the rows of **A** are iModulon activities, which represent the inferred TF activity in each sample. **(b)** A more detailed overview of the workflow, where entities in gray circles represent primary inputs (green text), secondary inputs (brown), or outputs (blue). The core ICA matrices are shown in blue boxes, and algorithms are shown in white boxes. Raw in-house or public data is processed to generate **X**, which is decomposed with ICA into **M** and **A**. The **M** columns are then thresholded and compared against known regulons when available, and **A** rows are interpreted and compared using differential iModulon activity (DiMA) analysis.

The iModulon workflow **(Figure 1.3a)** begins with a compendium of high quality transcriptomic expression data, **X**. ICA then decomposes this matrix into two matrices, the *m*odule matrix **M** and the *a*ctivity matrix **A**. Each iModulon has an associated column in **M**, representing its relationship to each gene, and a row in **A**, representing its activity in each sample. By thresholding the **M** matrix weights, we can obtain a set of genes, termed iModulon member genes, which are comparable to regulons. The **M** matrix therefore defines a quantitative structure for the TRN, an answer to problem (i) from **Section 1.2**. On the other hand, the **A** matrix summarizes the expression of all iModulon member genes in a single value for each sample. It represents a dimensionally reduced version of the original **X** expression data, which facilitates interpretation (question (ii) in **Section 1.2**). ICA is a matrix decomposition algorithm, so **X ~ M * A**. We can therefore use matrix multiplication and quantify the explained variance of iModulons, which in total typically falls between 60% and 90%. iModulon analysis is machine learning-based, which means that it improves with more data, meeting the need in problem (iii).

In the original iModulon paper [19], it was shown that iModulons are a highly informative framework for TRN investigation and summarization (See **Section 1.3.4**). Each iModulon represents an independent signal in the transcriptome, which can be categorized as regulatory, biological, genomic, or uncharacterized.

(i) Regulatory iModulons have gene sets which are statistically enriched for the regulons of known regulators, and they made up 61 of the 92 original *E. coli* iModulons. Note that regulons are defined by a variety of experimental methods [11], whereas iModulons are computed from transcriptomic data alone. This strong agreement lends credibility to the iModulon approach. The activity level of a regulatory iModulon is an inferred activity level for the regulator itself, which is very useful for gaining insights into the cellular state and sensory systems. The exact gene set of a regulatory iModulon may differ from the known regulon, which provides hypothetical relationships that can be further investigated and used to refine regulon annotations.

(ii) Biological iModulons have gene sets with similar functions, but no known shared regulator. This is an opportunity for discovery, as the shared function and apparent co-regulation indicates that at least one underlying regulator is likely present.

(iii) Genomic iModulons respond to changes to the genome that exist in some strains in the dataset. For instance, gene deletions appear as strong downregulation events in transcriptomes, and this is captured by iModulons.

(iv) Uncharacterized iModulons do not lend themselves to easy interpretation, and may be real transcriptional signals that require further inspection, or capture technical noise.

1.3.2 Independent Component Analysis

How does ICA identify these meaningful biological signals? Detailed discussions of the underlying machine learning are available [20], [22], [23], but a simplified description is provided here. We can imagine a dataset **X** as an *n*-dimensional cloud of samples, where *n* represents the number of genes on the order of thousands. Each column of **M** represents a direction in gene space, and each row in **A** represents the locations of each sample when projected onto that direction. For each component, ICA will begin with a random direction, project each sample onto it to obtain a possible **A** value for each sample, and then compute a measure of gaussianity or multi-information for the distribution of samples (in our case, the measure used is the 'log-cosh' function [24], [25]). After computing this value, ICA will then search nearby angles of rotation and also calculate their associated gaussianity. Because of the central limit theorem of statistics, it is predicted that noisy, non-biological signals will be gaussian, and the important iModulon signals will be non-gaussian. ICA therefore uses gradient descent to minimize gaussianity, and records each minimum as an iModulon.

Some readers may be familiar with principal component analysis (PCA), a similar but more common algorithm for matrix decomposition and dimensionality reduction [26]. The differences between the two methods can be useful for understanding ICA. Though PCA is typically computed with singular value decomposition, its results are equivalent to the following method: in the *n*-dimensional cloud of data, begin by choosing the direction with the maximum explained variance. Then, searching only through orthogonal directions to the first one, choose the next direction with the highest explained variance. Continue choosing orthogonal directions until all variance in the data is explained. The key differences are that PCA seeks to maximize explained variance whereas ICA seeks to minimize the gaussianity of each component, and PCA constrains each component to be orthogonal to all others.

PCA is useful for visualizing a dimensionally reduced dataset because of its high explained variance and orthogonality constraints. However, the components are typically dense, which would yield many member genes if passed through our thresholding method, and the genes in the components do not match well with known regulons. This is largely because the first component, with the highest explained variance, is a combination of many regulatory effects. The orthogonality constraint forces further components to also fail to match individual regulatory signals. Meanwhile, ICA has the flexibility to identify signals regardless of their explained variance, and the minimal gaussianity constraint is much more amenable to identifying the biological effects of interest. ICA's signals are also typically sparser (containing a fewer number of genes), which is more realistic for prokaryotic TRNs [27].

1.3.3 Assumptions and Limitations of ICA of Transcriptomes

ICA assumes that: (i) the dataset is composed of a linear combination of sources; (ii) sources are statistically independent from one another; and (iii) the distributions of source activities are not gaussian [23]. We now discuss how each one of these assumptions applies to TRN inference.

(i) Linear combination of sources: In this framework, each regulator or genomic alteration exerts control over its associated genes, and the sum total of all regulator-mediated changes defines the change to the TRN. The transcriptomes we analyze are pre-processed with the logarithm of transcripts per million, which preserves biological signals while enabling the use of linear methods [28], [29]. However, not all regulators will have a linear effect on their genes, and some regulatory interactions at complex promoters will violate the linear additivity of iModulon effects. We find in this dissertation that non-linear effects are often captured by multiple correlated iModulons (e.g. **Section 4.2.6.2**, **Figure 4.5c**), which still enables interpretation in low dimensions. However, it is important to be aware that complex regulation may lead to unusual artifacts in iModulon structures.

(ii) Sources are statistically independent: Unlike other TRN inference methods, ICA does not seek to directly describe a network in which targets of one regulator explicitly regulate another regulator, though real TRNs do exhibit that behavior. For instance, master regulators like RpoS in *E. coli* should be correlated in the **M** structure with downstream and interacting regulators, but the assumptions of ICA prevent this. Instead, we do observe meaningful correlations in the **A** matrix vectors. Other TRN inference methods that directly take the hierarchical regulatory structure into account require that modelers supply the algorithm with additional information about the interactions (e.g. [30]). This runs counter to being able to infer the TRN from data alone, which negates a major strength of this method.

(iii) Distributions of source activities are not gaussian: In other words, each iModulon must be differentially activated by some condition in the dataset. Take for instance the CecR regulon in *E. coli*, a set of genes that responds to antibiotics [31]. In an early version of PRECISE (PRECISE-124), the genes in this regulon were never perturbed by any condition, so their expression was normally distributed and they were not included in any iModulons. In a later version (PRECISE-278), a condition knocked out the regulatory gene *cecR*, which caused an upregulation of the CecR regulon genes in that condition. This regulon's expression distribution was now non-gaussian, and therefore the CecR iModulon was detected [19]. It is important to include many diverse conditions in the input data, which enables identification of more features in the TRN.

It is also worthwhile to note a few other limitations. iModulons do not capture all variance in the original datasets, which leaves ~10% – 40% of the variance as error. However, given ICA's ability to find consistent co-regulated signals, it is highly likely that the variance that is not explained is mostly noise or does not contain clear enough patterns to be interpreted. Also, ICA is stochastic, which led to the development of additional steps in the algorithm to ensure that the iModulons found are consistent across runs [19]. The algorithm is also sensitive to the choice of dimensionality, and a study developed an algorithm to tune this parameter [32]. Finally, it is difficult to threshold the **M** matrix for certain iModulons whose weight distributions are not bimodal, which results in some uncertainty about the genes which lie near the threshold.

In summary, iModulon interpretation is limited when transcriptional regulatory signals are non-linear, strongly interacting, or not activated in the underlying data. They do not capture all variance in the data, and parameter tuning is sometimes imperfect. Nonetheless, this framework has been shown to be highly informative, as the following section and remainder of this dissertation will demonstrate.

1.3.4 iModulons Show Promise as a Tool for Discovery

Prior to the work of this dissertation, we had significant evidence that iModulon analysis was a useful way to gain a transcriptome-wide summary of gene expression and reveal important new biological insights. Saelens et al. compared 42 similar methods of TRN inference, and found ICA to be the best at recovering known regulatory signals [17]. ICA was also applied by others to analyze a variety of yeast and human cancer datasets [33]–[37], and found to be the most robust factorization algorithm across a variety of datasets [38].

Sastry et al.'s analysis of PRECISE-278 revealed several important validations [19]. iModulons matched well with known regulons in 61 of 92 cases. The activity of 13 metabolic iModulons matched expectations under supplementation conditions with associated metabolites. The iModulons for CysB and MetJ matched TF binding sites better than previously defined regulons. The PurR regulon was bifurcated by iModulons, and the different subsets also had different promoter motifs. Two new high confidence regulons (YiaJ and YieP) were defined. The fear-greed tradeoff, which governs growth and stress responses (discussed further in **Section 4.2.8.1** and **5.2.3**), was quantified. Finally, the transcriptomic effects of mutations in various regulators were revealed. Taken together, the results showed that iModulons can provide a vast range of insights which can be validated in the laboratory.

In *E. coli*, iModulons were also used to study respiration [39], *oxyR* regulatory mutations [40], heterologous gene expression [41], two-component systems [42], and antibiotics [43]. A follow-up study on a gene that was unexpectedly included in one of the PurR iModulons revealed that there is a second regulator that responds to purine levels, named *punC* [44]. Importantly, the same pipeline was run on several different *E. coli* transcriptomic compendia, with similar iModulons resulting in each case. The findings of that study indicate that iModulons capture the real underlying TRN, and not dataset-specific patterns [45]. iModulons had a strong impact on our understanding of transcriptional regulation in *E. coli*, which motivated the work in this dissertation.

1.4 Aims of this Dissertation

iModulon analysis has been established as a useful tool in *E. coli*, but three goals needed to be met to facilitate wider adoption and application of this method. Firstly, it remained to be shown whether iModulon analysis would be as effective at describing the TRN in another model organism. *Bacillus subtilis,* an important gram-positive bacterium with an excellent candidate dataset, was selected for this purpose. Secondly, iModulon analysis is only useful to those who can view and analyze iModulons, which was difficult before the work of this dissertation. An online resource was needed to allow easy access to this data and knowledge. Thirdly, iModulon analysis reveals a new perspective on the TRN, which showed promise for analyzing the effects of changes to the genome. With prior methods, it was difficult to study the evolution of regulation, leaving systems-level adaptation mechanisms unclear. In order to better understand evolutionary mechanisms, it was important to integrate results from iModulons and the other levels of biology introduced in the beginning of the introduction (**Figure 1.1c**).

1.5 Dissertation Outline

We begin in **Chapter 2** by applying iModulon analysis to a transcriptomic compendium for *Bacillus subtilis*, which establishes that the workflow is applicable to a model gram-positive organism and reveals many interesting insights into the structure and function of the TRN, particularly with respect to the regulation of sporulation.

In **Chapter 3**, we develop iModulonDB, an interactive online knowledgebase which enables wide accessibility to iModulon structures. This platform makes iModulons findable, accessible, and reusable, which is necessary for their broad adoption as a major transcriptomic analysis method. As of the writing of the publication associated with Chapter 3, the knowledgebase contained three organisms (including *B. subtilis,* the results from the analysis in Chapter 2). We include an additional section (**Section 3.6**), which briefly summarizes the expansion of iModulonDB to include eight additional organisms in projects that I co-authored.

In **Chapters 4 and 5**, we apply iModulon analysis to interpret transcriptomic changes in laboratory evolution strains and combine it with other systems biology approaches. **Chapter 4** focuses on oxidative stress tolerance, while **Chapter 5** focuses on thermal stress associated with high temperatures. By connecting DNA mutations, RNA iModulon activities, and metabolic shifts, these chapters make iModulons interoperable. They reveal highly detailed mechanisms of stress tolerance, from a new perspective of global transcriptomic reallocation. The tolerance strategies are interesting for fundamental stress biology, and are also likely to influence biomanufacturing by providing design variables for the design of more stress-resistant cellular factories.

Taken together, this dissertation establishes iModulon analysis as a widely applicable, scalable, and interoperable method for understanding transcriptomic data. It reveals dozens of new insights about diverse biological systems, including sporulation, motility, stress tolerance, and redox metabolism. It stands upon decades of bottom-up characterization of cells and analyzes years' worth of accumulating datasets to reveal a new perspective on the sensory and regulatory systems of cells from the top-down. It reveals that a global understanding of the transcriptome in low dimensions is now both possible and useful as we build toward a more complete understanding of life.

Chapter 2. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome

The transcriptional regulatory network (TRN) of *Bacillus subtilis* coordinates cellular functions of fundamental interest, including metabolism, biofilm formation, and sporulation. Here, we use unsupervised machine learning to modularize the transcriptome and quantitatively describe regulatory activity under diverse conditions, creating an unbiased summary of gene expression. We obtain 83 independently modulated gene sets that explain most of the variance in expression and demonstrate that 76% of them represent the effects of known regulators. The TRN structure and its condition-dependent activity uncover putative or recently discovered roles for at least five regulons, such as a relationship between histidine utilization and quorum sensing. The TRN also facilitates quantification of population-level sporulation states. As this TRN covers the majority of the transcriptome and concisely characterizes the global expression state, it could inform research on nearly every aspect of transcriptional regulation in *B. subtilis*.

2.1 Background

Cells interpret dynamic environmental signals to govern gene expression through a complex transcriptional regulatory network (TRN). *Bacillus subtilis*, a model gram-positive soil and gut bacterium, is one of the most widely studied species in microbiology, providing a rich background for understanding its TRN. This generalist organism is a model for processes such as sporulation [46], biofilm formation [47], and competence [48] — all of which are key to understanding pathogenesis in other bacteria, such as *Staphylococcus aureus* and *Clostridium difficile*. *B. subtilis* is also commonly engineered for industrial production purposes [49], which creates demand for practical knowledge about how it responds to stimuli and alters its gene expression.

In 2012, Nicolas et al. [50] generated a transcriptomic microarray data set of *B. subtilis* with 269 expression profiles under 104 conditions, which included growth over time in various media, carbon source transitions, biofilms, swarming, various nutritional supplements, a variety of stressors, and a time course for sporulation. The wide scope and high quality of this data set have led to its broad adoption. It is now the expression compendium featured on *Subti*Wiki, an online resource for *B. subtilis* that is one of the most widely used and complete databases for any organism [12]. *Subti*Wiki contains detailed biological descriptions and binding sites for hundreds of transcriptional regulators; however, binding sites alone cannot explain the condition-specific transcriptomic responses of bacteria to dynamic environmental conditions [14], [15].

Independent component analysis (ICA) is an unsupervised statistical learning algorithm that was developed to isolate statistically independent voices from a collection of mixed signals [20]. ICA applied to transcriptomic matrices simultaneously computes independently modulated sets of genes (termed iModulons) and their corresponding activity levels in each experimental condition [19]. iModulons can be interpreted as data-driven regulons, though they rely on observed expression changes instead of transcription factor binding sites. The condition-dependent activity level of iModulons indicates how active the underlying regulator is. Since the number of iModulons is substantially fewer than the number of genes, they are a significantly easier way to analyze systems-level cell behavior.

ICA has been shown to extract biologically relevant transcriptional modules for a variety of transcriptomic datasets, especially in yeast and human cancer [33]–[37]. It was the best out of 42 methods at recovering known co-regulated gene modules in a comprehensive examination of TRN inference methods [17]. ICA also obtained the most robust modules across datasets compared to similar factorization algorithms [38]. We previously applied this approach to a large, high-quality *Escherichia coli* RNAseq compendium and extracted 92 iModulons, two-thirds of which exhibited high overlap with known regulons [19]. This analysis provided many insights into the *E. coli* TRN, including the addition of genes to known regulons (validated through chromatin immunoprecipitation), bifurcation of the purine synthesis regulon, the characterization of new regulons, and identification of clear associations between regulator mutations and activities. We have also applied ICA to transcriptomes of evolved strains to understand evolutionary trade-offs and regulatory adaptations in naphthoquinone-based aerobic respiration [39], and to characterize the function of the transcription factor OxyR, which responds to peroxide [40].

Without using ICA, others have attempted to infer the TRN of *B. subtilis*. Arietta-Ortiz et al. [30] used an "Inferelator" approach which utilized prior knowledge of the TRN along with transcriptomics (including the Nicolas et al. data) to obtain a global network, infer activity levels, and predict new TF-gene interactions. In addition, Fadda et al. [51] used genomic regulatory motifs of major regulators to infer a TRN, and Leyn et al. [52] combined a variety of available data types to infer regulons in *B. subtilis* as well as 10 related *Bacillales* species. These approaches have been valuable for expanding our understanding of the TRN and can be especially helpful in complex processes like sporulation where transcriptomics can be supplemented with other data types. However, prior methods suffer from a bias toward the known aspects of the TRN, which can pose a barrier for new discovery or unbiased validation of past data. They are also not as easily applicable to organisms with very incomplete TRN annotations. This motivates the development of fully unsupervised approaches like ICA.

Given our success with ICA applied to RNAseq data from a model gram-negative bacterium, we sought to determine what it can uncover about a microarray data set from a model gram-positive bacterium. Though RNAseq data exists for *B. subtilis*, the Nicolas et al. data set has a comparatively wider diversity of conditions and a more established reputation for data quality. We have shown that the condition space is more important than the technology used [19], [45], which makes this a good choice of data set. Using the wealth of TRN knowledge available on *Subti*Wiki, this analysis uncovers many insights. We determine the main functions and regulators that control a large fraction of the transcriptome, and we characterize the iModulon accuracy in relation to the known TRN. iModulon activities reveal relationships and stimuli that have been present in the data but never specifically investigated; it is therefore a powerful hypothesis-generating tool. We specifically present eight unexpected iModulon activations and hypotheses about their mechanisms. We characterize sporulation, which led us to the identification of three major transcriptomic stages in the process, including iModulons for the known sigma factor cascade. Finally, we present three transcriptional units with a little prior characterization that warrant further study.

2.2 Results

2.2.1 Independent Component Analysis Reveals the Structure of the *B. subtilis* Transcriptome

We performed ICA on the Nicolas et al. [50] data set (see **Section 2.4**) and obtained 83 robust iModulons. These 83 iModulons constitute the statistically independent gene expression signals found across the conditions used in the generation of this data. Together, they contain 36.25% of the genome and explain 72% of the variance in gene expression. The distribution of the number of genes in each iModulon follows a power law, similar to the power law for the connectivity of TFs in literature regulatory networks [53], [54].

Unlike regulons, which are sets of coregulated genes based on a variety of experimental results in the literature, iModulons are derived solely from the measured transcriptome through an unbiased method (**Figure 2.1a**). However, the known regulon structure of the TRN is largely recapitulated by the iModulons. 63 of the 83 iModulons were successfully mapped to a known regulator, and an additional 3 are likely to be co-regulated by unknown mechanisms. The iModulon-derived TRN covers 2235 gene/iModulon relationships, of which 1536 are known gene/regulator interactions and 699 are new. Our TRN structure contained seven iModulons that exhibited perfect overlap with annotated regulons and whose activity levels match expectations, such as MalR. This illustrates that independent signals such as transcription factor binding, which dictate gene expression, lead to observable signals in the TRN from condition to condition, and ICA was able to identify them. Graphical summaries of all iModulons, including their gene sets, activities, and overlap with regulons are presented online at iModulonDB.org under "*B. subtilis* Microarray" (See **Chapter 3**) [21].

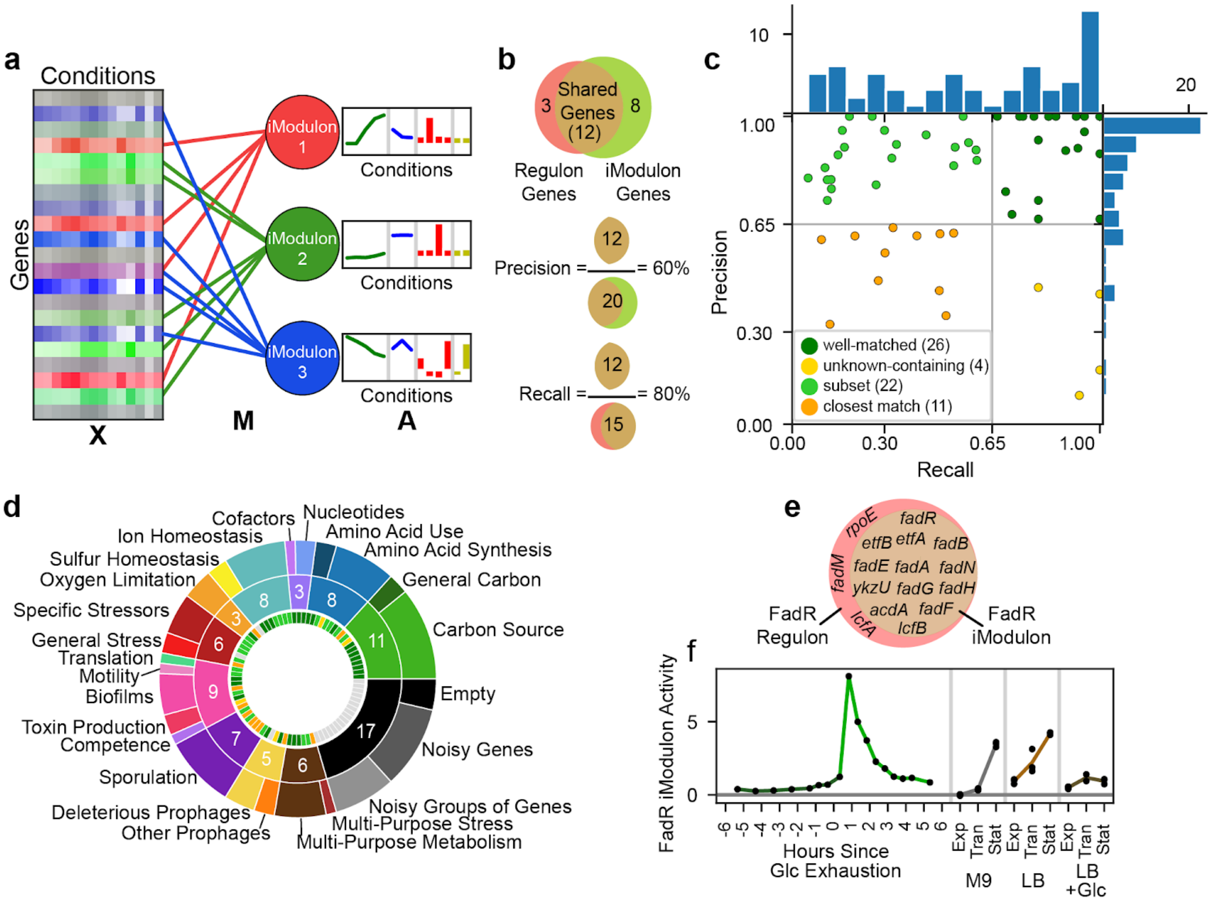


Figure 2.1: Overview of the iModulons of *B. subtilis.* **(a)** Given a matrix of gene expression data, **X**, ICA identifies independently modulated sets of genes (iModulons) in the transcriptome which are linked to genes through the matrix **M**. Three iModulons are symbolically represented; the red iModulon consists of four genes, and the green and blue iModulons consist of five genes. The condition-dependent activities of the iModulons are stored in matrix **A**. The bar chart indicates the activity levels of the iModulons under different conditions, where the colors indicate different experiments. The three matrices are related as **X** = **M*****A**. **(b)** Graphical representation of the definitions of precision and recall of a given iModulon and the corresponding regulon (example numbers are shown). **(c)** Scatter plot of precision and recall of the enrichments for the 63 (out of 83) iModulons that were matched to a regulon. Histograms in the margins demonstrate the high precision of most enrichments. **(d)** Donut chart of iModulon functions. The outermost ring lists specific functions and the center ring lists broad functions, with the number of iModulons in the broad category shown in white. The innermost ring shows the regulon confidence quadrant of the corresponding iModulon, as defined in c. **(e, f)** An example iModulon that was enriched for FadR. **(e)** Venn diagram of the FadR iModulon genes and the FadR regulon (non-coding RNAs have been omitted). **(f)** Activity level found in a row of **A** for four experiments (separated by vertical gray lines) from the data set. Activity levels increase during growth in the absence of glucose (M9 media, gray; LB media, light brown), remain low during growth in the presence of glucose (dark green, dark brown), and spike upon glucose (Glc) starvation (green). "Exp", "Tran" and "Stat" refer to exponential, transition, and stationary phase, respectively.

iModulons are given a short name, usually based on their enriched regulator. If multiple regulators control an iModulon, their names are separated by “+” to indicate the intersection of the regulons, or “/” to indicate the union of the regulons. In some cases, a different name was chosen based on the primary regulator, gene prefix, or most representative gene in the set.

The relationship between iModulons and regulators can be characterized by two measures: (1) precision (the fraction of iModulon genes captured by the enriched regulon) and (2) recall (the fraction of the regulon contained in the iModulon) (**Figure 2.1b**). These two measures can be used to classify iModulons into six groups (**Figure 2.1c**). (1) The well-matched group (n = 26) has precision and recall greater than 0.65. It includes several regulons with local regulators that are associated with specific metabolites. (2) The subset iModulons (n = 22) exhibit high precision and low recall. They contain only part of their enriched regulon, perhaps because the regulon is very large and only the genes with the most transcriptional changes are captured. This group contains global metabolic regulators such as CcpA and CodY, as well as the stress sigma factors. (3) A third group, deemed unknown-containing (n = 4), has low precision but high recall. These iModulons contain some co-regulated genes along with unannotated genes which may have as-yet-undiscovered relationships to the enriched regulators, or at least be co-stimulated by the conditions in the data set. (4) The remaining enriched iModulons are called the closest match (n = 11) because neither their precision nor recall met the cutoff, but the grouping had statistically significant enrichment levels and appropriate activity profiles. The difference in gene membership between these iModulons and their regulons provide excellent targets for discovery. The iModulons with no enrichments comprise the last two groups: (5) new regulons (n = 3) are likely to be real regulons with unexplored transcriptional mechanisms, while (6) the remaining uncharacterized iModulons were likely to be noise due to large variance within conditions or the fact that they contain one or fewer genes.

Functional categorization of iModulons provides a systems-level perspective on the transcriptome (**Figure 2.1d**). Metabolic needs account for approximately one-third of the iModulons, while comparatively fewer iModulons deal with stressors, lifestyle choices such as biofilm formation and sporulation, and mobile genetic elements like prophages. Some iModulons have multiple biological functions, such as one which synthesizes both nicotinamide and biotin. These iModulons may result from co-stimulation of the different functions by all conditions probed in the data set (e.g., both nicotinamide and biotin synthesis were always stimulated together by minimal media, so the algorithm could not separate them into unique signals).

The FadR iModulon provides an example of the information encoded by the iModulon gene membership (**Figure 2.1e**) and activities (**Figure 2.1f**). All genes within this iModulon are regulated by FadR, so this enrichment has 100% precision. Three genes that are annotated as belonging to the FadR regulon were not captured in the iModulon — *lcfA*, *rpoE*, and *fadM*. However, all three have additional regulation separate from that of FadR [55], [56], which may lead them to have a divergent expression from the rest of the iModulon. The activity levels (**Figure 2.1f**) reflect expectations: FadR genes are repressed by FadR in the presence of long-chain acyl-coA, and FadR itself is repressed by CcpA in the presence of fructose-1,6-bisphosphate [56], which causes the expression to rise as nutrients (specifically sugars and fats) are depleted, and to be particularly strong immediately following glucose exhaustion. As this example illustrates, the precision and recall are sensitive to developments in regulon annotations; they improve as regulon annotations become more complete [57].

2.2.2. iModulons Generate Hypotheses

iModulon activities can often be explained by prior knowledge, as was the case with FadR. However, they can also present surprising relationships that lead to the generation of hypotheses or strengthen arguments for recently proposed mechanisms. In the subsequent sections, we list eight such examples.

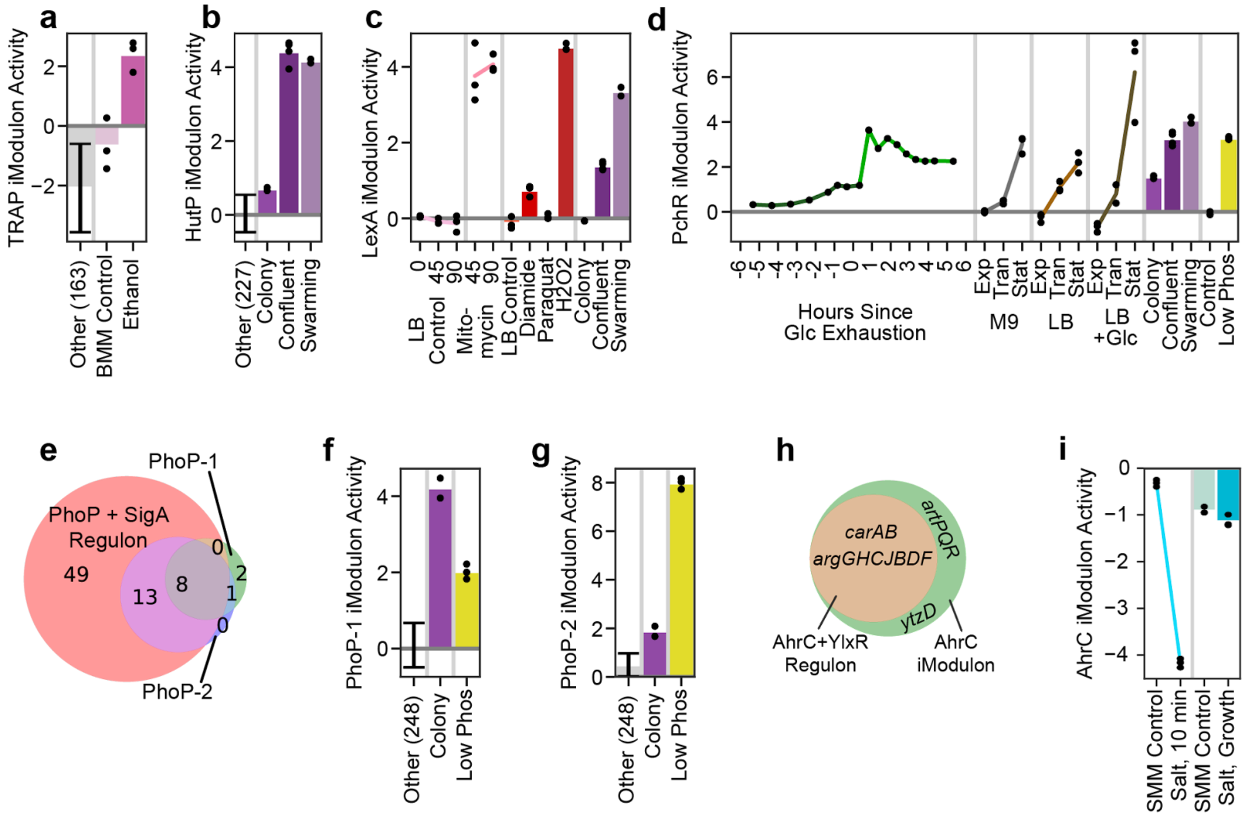


Figure 2.2: iModulons generate hypotheses. Error bars: mean ± standard deviation; black dots indicate separate samples; vertical gray lines separate different experiments in the data set. Unless otherwise stated, "Other" category includes all conditions except sporulation and those shown, with the number of samples included in parentheses. **(a)** Tryptophan synthesis (TRAP) iModulon activity, which is unexpectedly elevated by ethanol. The experiment was carried out in Belitsky minimal medium (BMM). The "Other" category excludes carbon source transition experiments, in which this iModulon exhibits technical noise. **(b)** Histidine utilization (HutP) iModulon activity, which is strongest in quorum conditions. **(c)** LexA iModulon activity is elevated by DNA damage (mitomycin and peroxide) and in swarming. **(d)** Pulcherrimin (PchR) iModulon activity increases when growth is expected to slow, especially in the stationary phase in rich (LB) media containing glucose (Glc). "Exp", "Tran" and "Stat" refer to exponential, transition, and stationary phase, respectively. **(e)** Venn diagram of gene presence in the PhoP+SigA regulon and related iModulons. Numbers indicate the amount of genes or non-coding RNAs in each subset. Although the iModulons are significantly enriched for the intersection of the PhoP and SigA regulons, they have been named PhoP-1 and PhoP-2 for simplicity. **(f, g)** Bar graphs of PhoP iModulon activity demonstrating the use of PhoP-1 for early biofilm growth ("Colony" refers to individual colonies on a plate after 16 h) and PhoP-2 for extreme phosphate starvation ("Low Phos" indicates phosphate starvation for 3 h). **(h, i)** The arginine synthesis (AhrC) iModulon. **(h)** Venn diagram of the arginine synthesis (AhrC) iModulon and regulon; the regulon contains additional arginine-related genes (*artPQR* and *ytzD*) that are not known to be regulated by AhrC. **(i)** AhrC iModulon activity in osmotic stress conditions. This iModulon is surprisingly downregulated by salt shock. N = 3 samples for all conditions shown.

2.2.2.1 Ethanol May Stimulate Tryptophan Synthesis

The tryptophan synthesis iModulon (*trpEDCFB*) was strongly activated under ethanol stress

(**Figure 2.2a**), a response that has not been previously documented in bacteria. This iModulon is regulated by the *trp* attenuation protein (TRAP), which represses its genes in the presence of tryptophan [58]. Therefore, this activation indicates that ethanol is probably depleting intracellular tryptophan concentrations. Exploring the tryptophan synthesis pathway reveals a hypothetical mechanism for this depletion: flux from the precursor chorismate may be redirected to replenish folate that has been damaged by ethanol oxidation byproducts [59]. If this hypothesis is accurate, it may inform research on the tryptophan deficiency and neurotransmitter metabolism problems observed in human alcoholic patients [60], [61], especially given that *B. subtilis* is an important folate producer in the gut microbiome [62], [63].

2.2.2.2 Histidine May Be Utilized by Quorums

The HutP iModulon for histidine utilization (*hutHUIGM*) is controlled by an antiterminator that derepresses it in the presence of excess histidine, as well as by the master regulators CcpA and CodY; therefore, its activation indicates that histidine is plentiful while other amino acids are not and that carbon sources are poor [64]. Surprisingly, it was by far most strongly activated in confluent biofilms and swarming cells (**Figure 2.2b**). Independent colonies from the same experiment do not exhibit activation, which leads us to rule out the media composition as the reason for these activity levels. The connection between these lifestyle conditions and histidine metabolism has not been studied in *B. subtilis*, but it has been observed in *A. baumannii*, where histidine degradation was shown to be upregulated in proteomic studies of biofilms, and histidine supplementation stimulated increased biofilm production [65]. Two recent studies discovered that biofilm-inhibiting antimicrobials worked by suppressing histidine synthesis in *Staphylococcus xylosus* [66], [67]. One proposed mechanism implicated the production of extracellular DNA, which is an important component of both *A. baumannii* and *B. subtilis* biofilms [68]. Given that this iModulon is also activated by swarming cells, an alternative hypothesis may be that HutP is involved with quorum sensing or surfactant production: both activating conditions have a quorum and high surfactant production, while independent colonies do not.

2.2.2.3 DNA Damage May Stimulate Swarming

The LexA iModulon regulates the SOS response for DNA protection and repair. It is strongly activated by three conditions (**Figure 2.2c**). LexA stimulation by mitomycin and hydrogen peroxide is expected since those conditions damage DNA [69], [70]. Unexpectedly, this iModulon is also activated in swarming cells despite a lack of DNA damaging agents in that condition. We propose a potential mechanism for this activation: recent research has indicated that certain cells in a culture will tend to accumulate reactive oxygen species and DNA damage. Those cells will produce Sda (a developmental checkpoint protein) and form a subpopulation separate from those that produce biofilm [71]. The LexA+, biofilm− population would no longer be producing EpsE, which catalyzes a step in the biofilm synthesis process and also suppresses swarming [72]. In addition, this connection may be mediated by interactions between RecA and CheW, which have been observed in *Salmonella enterica* [73]. Therefore, we predict that DNA damage encourages swarming motility based on iModulon activation and this mechanism.

2.2.2.4 An Iron Chelator May Signal the Stationary Phase

The PchR iModulon produces, extrudes, and imports pulcherrimin, an iron chelator [74]. Over all of the exponential to stationary phase growth experiments, we observe increases in PchR activation (**Figure 2.2d**). We also see PchR activation in late-stage biofilm, glucose exhaustion, and phosphate starvation experiments. These results agree with a recent study that found pulcherrimin to be an important intercellular signal for the stationary phase that also helps exclude competing bacteria from established biofilms [75]. The regulation mechanisms of iModulons like this one can be the subject of future research.

2.2.2.5 Phosphate Limitation Stimulates Tiers of Regulation

The PhoP regulon controls phosphate homeostasis. It appears as two separate iModulons (**Figure 2.2e–g**). PhoP-1 encodes high-affinity phosphate uptake transporters. Phosphate is used to produce (and is effectively stored in) teichoic acid, which is a major component of the cell wall. As a colony grows, it must uptake phosphate to produce more cell walls — indeed, teichoic acid intermediates are the major stimulus for PhoP activity [76]. It is therefore unsurprising that PhoP-1 is strongly activated in independent colonies, which are exponentially growing in close quarters with low local free phosphate concentrations. PhoP-2 contains PhoP-1 as well as 13 other genes which encode more extreme phosphate recovery strategies: *phoABD*, which salvages phosphate monoesters but produces reactive alcohols, *glpQ*, which degrades extracellular teichoic acid, and *tuaBCDEFGH*, which replaces teichoic acid with phosphate-free teichuronic acid. PhoP-2 is only active under phosphate starvation, consistent with the extreme strategy it encodes. Perhaps the affinities of the promoters of the PhoP-2 specific genes are lower than that of the PhoP-1 genes, which could lead to this graded response.

2.2.2.6 The Arginine Synthesis iModulon Provides Two Important Insights

The arginine synthesis iModulon provides two interesting insights. Its genes include *argGHCJBDF* and *carAB*, which are known to be repressed in the presence of arginine by AhrC [77]. The first insight is that it also contains *artPQR* (**Figure 2.2h**), which are arginine importers not known to be transcriptionally regulated by AhrC – given that they are part of the same independent signal in the transcriptome, they likely share this regulation.

In addition, the iModulon was unexpectedly downregulated in salt shock, but not after growth in salt (**Figure 2.2i**); this has not been explored in previous studies. Here, the putative mechanism is less clear. It may involve the production of osmoprotective solutes such as proline [78], which might perturb metabolic networks in such a way that arginine concentrations increase and then downregulate these genes. After proline stores have been established, arginine concentrations appear to be restored. There is also evidence of a proline/arginine metabolic link in another iModulon: the RocR/PutR iModulon combines the utilization of both amino acids into one signal. Exploration of this relationship may help to understand broader changes in amino acid metabolism and its regulation under stress conditions.

2.2.2.7 The CcpA Regulon Is Captured By Two iModulons

The CcpA iModulons regulate carbon catabolites in different phases of growth (**Figure A.1**), which may suggest divergent preferences for carbon catabolites determined by growth phase and starvation state; the same catabolites that are preferred during exponential growth are preferred during germination. CcpA-1 contains mostly sugar metabolism enzymes (ribose, sucrose, mannose, trehalose, lichenan, etc.) while CcpA-2 contains a mix of genes including those for inositol consumption, tricarboxylic acid permeability, and acetyl-CoA utilization. CcpA-1 thus represents the relatively more preferred alternative carbon sources.

2.2.2.8 Correlations Between iModulon Activity and Regulator Expression Reveal Mode of Regulation

Regulatory proteins are often subject to ligand binding or kinase activity, which switches them between active and inactive states. Therefore, the gene expression of a given transcription factor does not usually correlate with the expression of its targets; this is the case with MalR, which is activated through phosphorylation by MalK only when malate is present [79]. Since iModulons combine co-regulated gene expression into easy-to-evaluate activity levels, we can attempt to correlate regulator expression with iModulon activity. As expected, many of these correlations are low. For example, MalR activation occurs through a post-transcriptional binary switch, so there is no correlation between MalR gene transcription and iModulon activity (**Figure A.2a**). A major exception to this is the sigma factors, which are often only regulated at the expression level. In these cases, we observe much higher correlations between expression and activity, such as with the motility sigma factor, SigD (**Figure A.2b**). High correlations are also observed when the regulator undergoes positive feedback, in which case it is a member of its own iModulon (**Figure A.2c**).

The Thi-box is a riboswitch that is conserved in all domains of life and regulates the expression of genes for thiamine synthesis and transport. In *B. subtilis*, this sequence is upstream of a transcriptional terminator, which it deactivates in the absence of thiamine [80], [81]. We would therefore expect the Thi-box sequence to be constitutively expressed, and its downstream genes to respond to thiamine levels – there would be no correlation between Thi-box RNA expression and the activity levels of its genes. Instead, this relationship had a unique shape which was consistent for all 5 Thi-boxes (**Figure A.2d**). We believe that this may be explained by differential degradation of the short Thi-box RNA sequence. Under minimal media conditions, the thi-box sequence does not bind thiamine, so RNA polymerase reads through it and produces a long, relatively stable RNA molecule, which is measured as both high Thi-box expression and high Thi-box iModulon activity. Under rich media conditions, the binding of thiamine terminates transcription, preventing Thi-box iModulon activity and producing a short, less stable RNA molecule. Interestingly, this short sequence may be degraded quickly in flasks (evidenced by a lack of apparent Thi-box expression) but appears to remain in biofilms for long enough that it could be measured in this experiment. Little is known about differential RNA degradation in biofilms, but this result motivates further study of that phenomenon.

2.2.3 Six iModulons Capture the Major Transcriptional Steps of Sporulation

The data set we analyzed contained an eight-hour sporulation time course, which yielded six major sporulation iModulons that were activated sequentially over the first 6 hours (**Figure 2.3a**). The identification of these gene sets by ICA indicates coherent expression across the transcriptome, and more dramatic transcriptional variation compared to excluded genes. The conclusions drawn from these iModulons are limited by the complexity of sporulation [46], [82] and the stochasticity of its onset [83]. Because of this, we observe many genes shared between consecutive iModulons. Nonetheless, the following analysis demonstrates that they still provide valuable information, including identifying 20 uncharacterized proteins whose annotations did not previously reflect a putative relationship to sporulation.

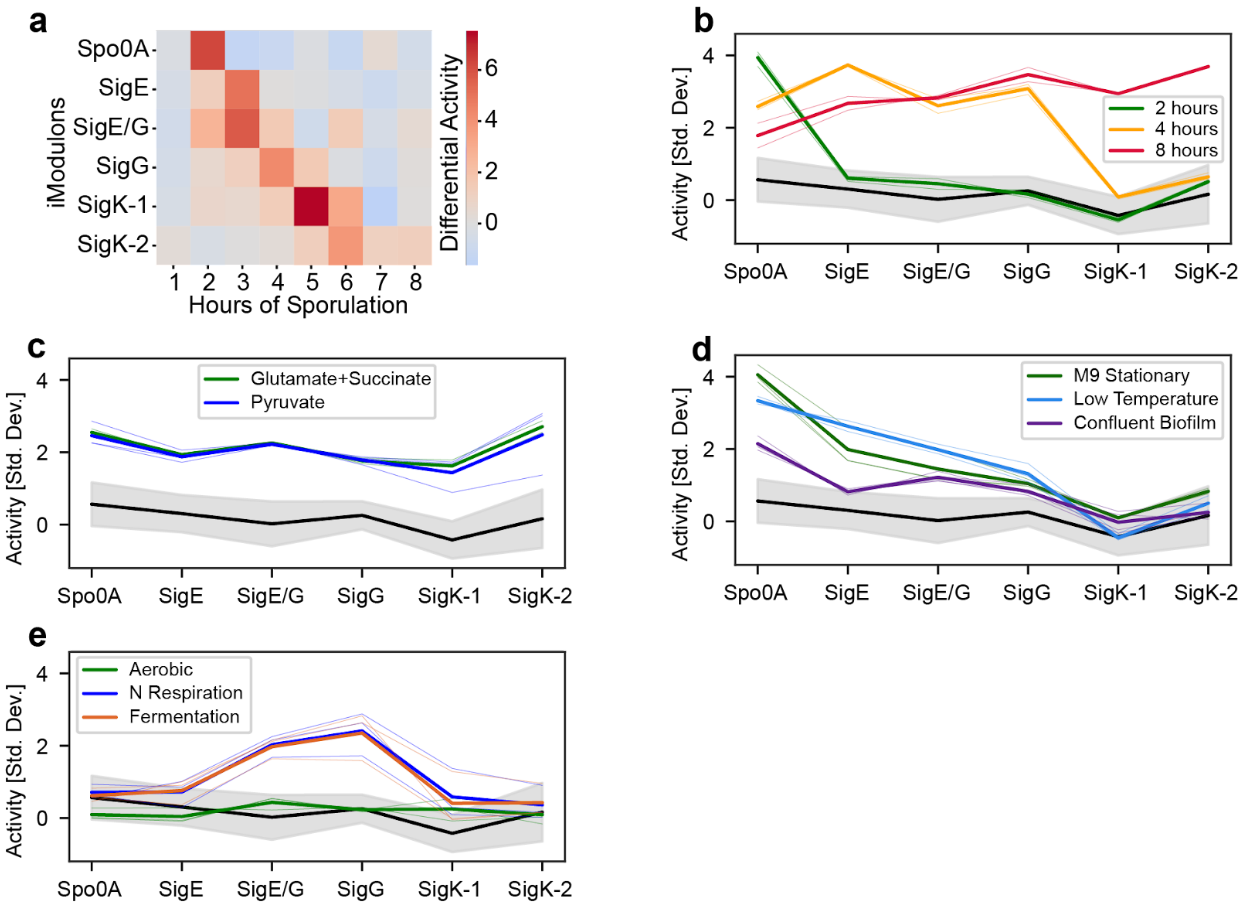


Figure 2.3: Sporulation iModulons reveal the tendency of certain conditions to sporulate. **(a)** Heatmap color indicates the change in iModulon activity over the previous hour. **(b–e)** Line plots of the sporulation progression for selected conditions, with thick lines indicating mean activity and thin lines indicating individual samples. Activity levels were divided by the standard deviation. The black line surrounded by a shaded gray region is the average of all conditions not shown in any plot ± standard deviation (n = 200 samples). **(b)** Three time points of sporulation, showing Spo0A activation at sporulation onset (2 h, green), cumulative expression up to the fourth step (SigG) for an intermediate time point (4 h, orange), and expression of all stages at 8 h (red). **(c)** Minimal media supplemented with these carbon sources leads to expression of all sporulation iModulons. **(d)** Three conditions reached the intermediate steps of sporulation. **(e)** Anaerobic conditions exhibit unusual activity. "Aerobic" is the control condition.

The gene sets and regulators of the sporulation iModulons roughly match the known sporulation progression. The Spo0A iModulon contains mostly genes known to be activated by high levels of Spo0A~P, including the sigma factors for upcoming sporulation steps, chromosome preparation machinery, and septal wall formation. It is rapidly activated between hours 1 and 2 of the time course (**Figure 2.3a**). Next, the SigE iModulon carries out functions in the mother cell for engulfment of the forespore. After SigE, a dual SigE/G iModulon is activated, which regulates early spore coat formation by both the mother and forespore cells. The SigG iModulon follows; it contains germination receptors, metabolic enzymes, and stress resistance genes. Finally, the SigK regulon is split into two iModulons with functions including coat maturation and mother cell lysis. The difference between the two SigK iModulons may partially be explained by the action of the TF GerE, which represses members of SigK-1 and activates a large fraction of SigK-2. This is consistent with the known temporal regulation of the SigK regulon [84]. Notably, SigF is the only absent sigma factor; we believe it was not identified because its genes are expressed simultaneously with the SigE, SigG, and SigE/G iModulons, and because many SigF genes are also under SigG control [85]. Nonetheless, these functions and regulators largely match expectations based on literature, providing an *a priori* validation of the set of known sporulation steps.

The activity levels of the sporulation iModulons can be viewed as markers of progress through sporulation: high Spo0A activity indicates that new spores are forming, and high SigK-2 indicates that some spores are completing the process. Therefore, we can understand how far along other conditions are based on their sporulation activity levels (**Figure 2.3c–e**). Most conditions have a very low level of activation, but the "glutamate + succinate" and pyruvate supplements to minimal media conditions both have elevated expression across all sporulation iModulons, which indicates that the poor carbon sources in these conditions stimulated sporulation (**Figure 2.3c**). Indeed, pyruvate has been shown to regulate sporulation [86], [87]. Some other conditions appear to have made it partway through the process: confluent biofilms, the stationary phase in minimal media, and growth at cold temperature all reached the third of six steps. This is appropriate for these conditions based on previous studies [88]–[90] (**Figure 2.3d**).

With one exception, the progression from one sporulation iModulon to the next is cumulative: we do not see strong activation of step 2 unless step 1 is active, and so on. This agrees with prior observations [91]. The only exception to this rule is elevated SigG activity by cells in anaerobic conditions (**Figure 2.3e**). This connection is also evident from gene presence: a flavohemoglobin required for anaerobic growth, *hmp*, is part of the iModulon despite no known connection to SigG.

Previous studies have also acknowledged that some SigG-dependent genes are required for anaerobic survival [92]. However, it is known that ectopic activation of SigG is limited by negative feedback [93], [94] and unlikely to occur in vegetative cells [95]. We, therefore, propose further experiments to determine the role of SigG-dependent genes in anaerobiosis.

2.2.4 Changes in iModulon Activity Reveal Global Transcriptional Shifts During Sporulation

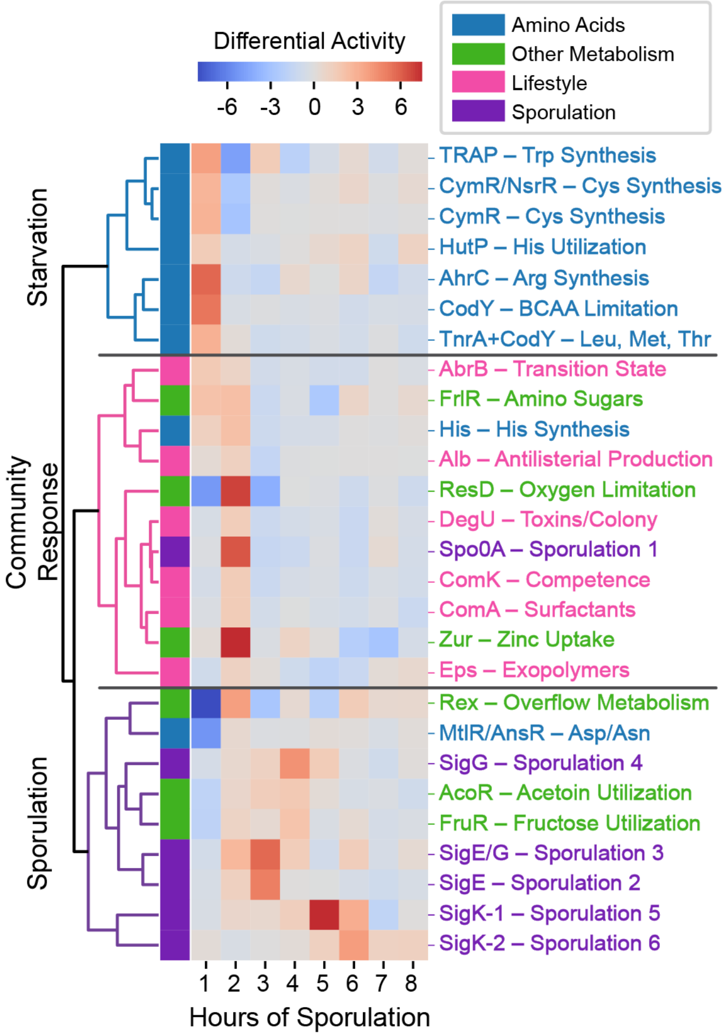


Figure 2.4: Global reallocation during sporulation. Heatmap color indicates a change in iModulon activity over the previous hour. Selected iModulons were hierarchically clustered according to the Pearson R correlation between sporulation activity derivatives.

In complex processes such as sporulation, the entire cellular transcriptome undergoes system-wide changes beyond those directly related to the process at hand. While much effort has been put into understanding metabolic changes at the onset of sporulation [46], [89], [91], metabolic and lifestyle-related regulatory activity are difficult to summarize concisely with previous methods. Because ICA provides a simple method for tracking transcriptome-wide changes, we analyzed activity level fluctuations for the sporulation time course (**Figure 2.4**). Three major stages are involved: a self-preserving metabolic response to amino acid starvation in the first hour, a community-wide lifestyle reallocation in the second hour, and progression through sporulation in the remaining time points.

In the first hour, many amino acid synthesis iModulons (tryptophan, cysteine, arginine, leucine, and threonine) and one amino acid utilization iModulon (histidine) are rapidly activated. This is likely the result of amino acid starvation by the sporulation media, which derepresses these iModulons through TFs including CodY. CodY also derepresses the fructosamine consumption iModulon [96] at this time. The AbrB iModulon is derepressed; it responds to nutrient limitation through a variety of functions, including cannibalism [97], that herald the stationary phase and prolong entry into sporulation.

In the second hour, Spo0A is strongly activated in a process that has been widely studied; this marks the onset of sporulation [98]. Also, the histidine utilization of the first hour is compensated by histidine synthesis in the second hour. Zinc, an important cofactor for sporulation proteins [99], [100], is taken up. Various colony, biofilm, and antimicrobial iModulons are activated to support the forming spores (DegU, ComA, Eps, Alb). ComK, the competence iModulon, is expressed as an alternative response to starvation. ComK's brief activation at this time point is consistent with the short competence window observed before commitment to sporulation [48]. We also observe the activation of ResD, which is typically associated with anaerobic conditions [101], [102], and Rex, which regulates overflow metabolism, providing interesting connections to the potential anaerobic activity of SigG discussed in the previous section.

As sporulation continues, fewer non-sporulation iModulons are activated. The notable exceptions are AcoR and FruR, which are both activated around the fourth hour. Both acetoin and polymeric fructose function as extracellular energy stores [103], [104], so perhaps they are used at this stage to provide a final energy source for the completion of sporulation. Overall, these results demonstrate an application of ICA for observing transcriptome-wide changes and lay out the major population dynamics and metabolic changes that underscore spore formation.

2.2.5 Some Poorly Characterized iModulons May Perform Important Functions

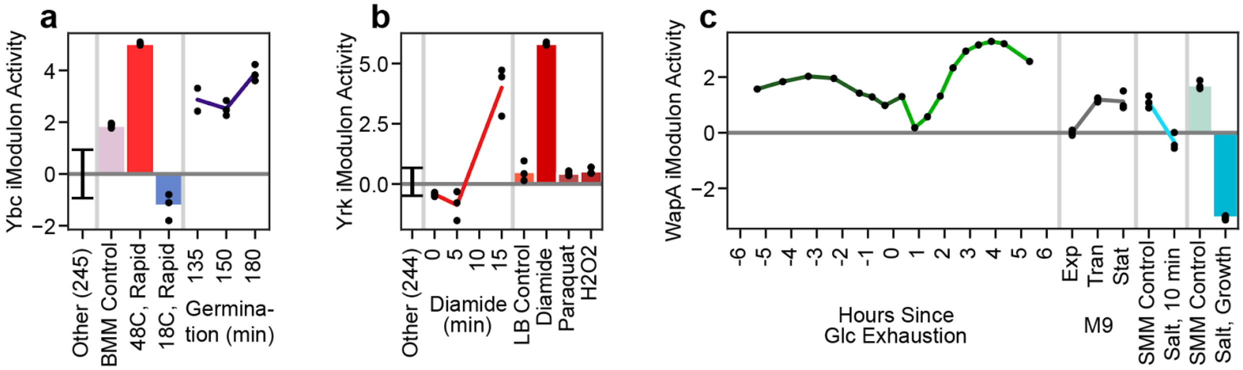


Figure 2.5: The activity levels of uncharacterized iModulons agree with their putative functions. Bars and lines indicate means, black dots indicate individual samples, and error bars indicate one standard deviation. The “Other” category includes all conditions except the ones in the plot, with the number of samples included in parentheses. Vertical gray lines separate different experiments in the data set. **(a)** The activity levels of the Ybc iModulon indicate that it may be a response to heat shock or germination. The Belitsky minimal media (BMM) control occurs at 37 °C. **(b)** The activity levels of the Yrk iModulon (putative sulfur carriers) suggest that it is a response to diamide. The three conditions on the right were taken from LB cultures 10 min after exposure to the labeled stressor. **(c)** The activity levels of the WapA iModulon indicate activation by nutrient limitation (glucose exhaustion and the three growth phases of M9 media) and suppression by osmotic stress, both in the short (light blue time course) and long term (bars). “Exp”, “Tran”, and “Stat” refer to exponential, transition, and stationary phase, respectively.

Given the vast number of uncharacterized genes in bacterial genomes, ICA can help to narrow the search for new and important regulons by identifying groups of genes with transcriptional coregulation and their corresponding activity levels. We have identified three iModulons that warrant further study. The first, the *ndhF-ybcCFHI* operon, may be involved in heat shock and germination (**Figure 2.5a**). Another, the *yrkEFHI* operon, contains putative sulfur carriers that are very likely to assist in the cellular response to diamide stress (**Figure 2.5b**). Finally, the WapA iModulon reveals additional genes which may contribute to an interspecies competition strategy (**Figure 2.5c**). The other uncharacterized iModulons which are not likely to be noise are prophage elements, whose regulatory mechanisms and effect on phenotype warrant further study.

The *ndhF-ybcCFHI* operon was identified as its own iModulon (**Figure 2.5a**). *ndhF* is known to be a subunit of NADH dehydrogenase, but the *ybc* genes have not been characterized at all. Peptide homology suggests that *ybcC* may form a protein that binds to *ndhF*, and that *ybcF* may be carbonic anhydrase. The activity levels of this group of genes demonstrate very strong activation under heat shock, as well as repression during cold shock and unusually high germination activity. Perhaps this is a new category of heat-responsive genes; since heat shock is a complex response [91], a small operon like this may have been overlooked in previous studies. We can hypothesize mechanisms through which this operon might benefit the cell: heat shock should upregulate *ndhF* to help to power the heat stress response, *ybcC* might be a chaperone for *ndhF*, and maybe *ybcF* assists in raising the pH after a temperature increase lowers it. We propose gene knockout experiments to validate that these genes play a role in the survival of heat shock.

Another uncharacterized iModulon is the *yrkEFHI* operon. None of these genes have been characterized, but two of them are putative sulfur carriers, and one is a putative sulfurtransferase. This iModulon exhibits consistently low activity except in the two conditions with ten or fifteen minutes of diamide exposure (**Figure 2.5b**). Since diamide oxidizes thiols to disulfides, it would make sense for sulfur carriers to be necessary in this condition. Future experiments can be performed to confirm this reasoning and identify transcriptional regulatory mechanisms.

Also, the WapA iModulon contains several uncharacterized genes that may be coregulated by YvrHb, DegU, and WalR and participate in a unique, recently discovered interspecies competition mechanism [105]. This system protrudes fibers from the cell wall to deliver the WapA tRNase to enemy bacteria, potentially compromising cell wall integrity for greater nutrient availability. We observe activation of this iModulon under starvation conditions and repression under cell wall stress (**Figure 2.5c**), consistent with its putative function.

2.3 Discussion

Here, we decomposed the existing, high-quality *B. subtilis* expression data set [50] using ICA. This decomposition identified 83 iModulons in the transcriptome whose overall activity can explain 72% of the variance in gene expression across the wide variety of conditions used to generate the data set. Sixty-six of the iModulons correspond to specific biological functions or transcriptional regulators. We analyzed the gene sets and activity levels of the iModulons and presented findings that either agree with existing knowledge or generate hypotheses that could be tested in future studies. The remaining 17 iModulons are independent signals with no coherent biological meaning.

Through the application of ICA, we were able to identify well-studied gene sets with high accuracy (such as the MalR and FadR iModulons), and uncover insights that suggest candidate underlying mechanisms. We discovered unexpected relationships between stress, metabolism, and lifestyle: ethanol appears to stimulate tryptophan synthesis, histidine utilization may be a feature of quorum sensing, DNA damage may induce swarming, and the iron chelator pulcherrimin could help to signal the stationary phase. The tiered response to phosphate limitation was captured as two separate iModulons, which may provide evidence for variable promoter affinity across the known regulon. ICA accurately decomposed sporulation into a small set of steps which allow sporulation progress to be tracked; this revealed unexplained, unusual activity for SigG in anaerobic conditions. The global transcriptional response to sporulation in metabolism and lifestyle governance was summarized concisely in three stages by iModulon activities. Finally, three iModulons contain mostly uncharacterized gene sets, which represent a promising area for further research. Overall, we have demonstrated that ICA produces biologically relevant iModulons with hypothesis-generating capability from microarray data in this model gram-positive organism.

The iModulon genes and activity profile data, along with graphical summaries are available for examination by microbiologists with specific interests about functions in *B. subtilis* that are not detailed in this dissertation. We also have an online resource, iModulonDB.org (**Chapter 3**), where users can search and browse all iModulons from this data set and view them with interactive dashboards [21]. Code for the analysis pipeline used here is available on github (https://github.com/SBRG/precise-db). There is a strong potential for protein identification, transcription factor discovery, metabolic network insights, function assignment, and mechanism elucidation derived from this iModulon structure of the TRN.

As with all machine learning approaches, the results from ICA improve as it is provided with more high-quality data [19]. Future research may append unique conditions to this data set and observe the changes to the set of iModulons it finds. Perhaps multi-purpose iModulons will be divided into their biologically accurate building blocks, the noise will be removed, and new regulons will emerge as the signal-to-noise ratio improves. With enough additional data, ICA could potentially characterize the entire TRN in great detail, a goal that has been the subject of research for over half a century. Ultimately, this could be the foundation for a comprehensive, quantitative, irreducible TRN.

2.4 Methods

2.4.1 Data Acquisition and Preprocessing

We obtained normalized, log2-transformed tiling microarray expression values from Nicolas et al. [50] (GEO accession number GSE27219), which span 5875 transcribed regions (4292 coding sequences and 1632 previously unannotated RNAs) and 269 sample profiles (104 conditions). The strain used, BSB1, is a prototrophic derivative of the popular laboratory strain, 168. Three samples (S3_3, G + S_1, and Mt0_2) were removed so that the Pearson R correlation between biological replicates was not < 0.9, except in the case of sporulation hour 8, where n = 2 and R = 0.89. To obtain more easily interpretable activity levels, we centered the data by subtracting the mean in the M9 exponential growth condition from all gene values. This is consistent with our prior work in *E. coli*, where a similar condition was chosen for this purpose. All activities are therefore relative to a known, consistent baseline condition.

2.4.2 Independent Component Analysis

Independent component analysis decomposes a transcriptomic matrix (**X**) into independent components (**M**) and their condition-specific activities (**A**): **X** = **M** ∗ **A**. Note that the **M** matrix was previously called **S** [19]; it has been changed to avoid confusion with other nomenclature.

We processed the quality-checked, centered data (**X**) with the Scikit-Learn (v0.19.0) implementation of FastICA [24] using 100 iterations, a convergence tolerance of 10^{-7}, log(cosh(x)) as the contrast function, and parallel search. We calculated enough components to reconstruct 99% of the variance as determined by PCA.

After 100 iterations of ICA, the **M** matrices were pooled and clustered with Scikit-Learn DBSCAN [24] (epsilon = 0.1, minimum size = 50) in order to find robust components which appear in each random restart. Since identical components can have opposite signs, we defined distance for this algorithm using a sign-agnostic method:

$$d_{x,y} = 1 - |\rho_{x,y}|$$

where $d_{x,y}$ is the distance and $\rho_{x,y}$ is the Pearson correlation between components x and y. Components belong to a cluster if $d_{x,y}<0.1$ with all other components in the cluster. To ensure repeatability, all signs in a cluster were inverted if necessary so that the highest weighted gene would have a positive sign. The centroids of each cluster defined the weightings in M and were used to calculate A.

This process was repeated 100 times (for a total of 10,000 ICA runs), and components that did not arise in every run were discarded. The result contained 83 robust components.

We normalized each component in the **M** matrix such that the maximum absolute gene weight was 1. We performed the inverse normalization on the **A** matrix to conserve the same values. Therefore, each unit in **A** is equivalent to a unit log change in expression if the iModulon were to contain only one gene.

2.4.3 iModulon Threshold Determination

The distribution of **M** matrix weights of each gene for a given component consists of a large number of near-zero values along with a small number of genes at the tails. To define the gene set of the iModulon, we need to choose a threshold value that separates the normally distributed near-zero genes from the more meaningful, non-gaussian tails. To do so, we used Scikit Learn's implementation of the D'Agostino K2 test, which quantifies the skew and kurtosis of the distribution as a measure of gaussianity [24], [106]. We iteratively remove the gene with the highest absolute value in the component and calculate the K2 value until the value falls below a K2 cutoff value (1300). All genes that were removed are members of the iModulon gene set, and the non-removed genes are sufficiently normally distributed around zero to be considered noise. In all cases discussed in this chapter, all member genes have positive weights, which allows for easier representation as a set of genes and a simple interpretation of activity.

The cutoff value of 1300 was determined by a sensitivity analysis. Over a range of cutoffs (200 - 2200), we computed the top regulator enrichments and F1 scores as described in **Section 2.4.4**. The cutoff with the highest mean F1 score was selected. In seven cases, this cutoff was not appropriate because it removed all genes from the iModulon (5/7) or captured non-important genes (2/7), so the threshold was adjusted to 500 or increased slightly as necessary. In the two cases (MalR and Rex) in which the threshold was increased, we observed the tails of the distribution starting a bit higher than their computed thresholds, and the full tail corresponds to a regulon. For this reason, it was appropriate to raise the thresholds in those cases.

2.4.4 Regulator Enrichment

Regulon information was obtained from *Subti*Wiki [12]. For each iModulon, we obtained all regulators that regulate any gene in their gene sets. We also used all combinations of regulators, denoted by "+" between regulator names, to capture regulons with more than one regulator. For each of those individual regulators and regulator combinations, we obtained a regulon set, a list of all genes that share that regulation. Next, we computed p-values for each regulon's overlap with the iModulon gene set using the two-sided Fisher's exact test (FDR < 10^{-5}) [24], [107]. We also computed F1 scores, which are the harmonic averages of precision and recall.

After the sensitivity analysis (**Section 2.4.3**) determined the appropriate cutoff, significant enrichments for each iModulon were then manually curated. In most cases, the most significant enrichment was chosen. Some iModulons appeared to be a combination of two or more significantly enriched regulons, so their assigned regulator was a union of both, denoted by "/" between regulator names.

Our regulator enrichments have very high precision and recall scores, but they have an inherent bias because the threshold for iModulon membership was chosen to maximize them. Our method of selecting the threshold improves with the completeness of the TRN annotations, and would be ineffective for an organism with a very incomplete TRN. We could work around that limitation with approaches using other gene groupings, such as functional, category, or motif enrichments, or by developing approaches that compare iModulons across organisms, such as comparing iModulon size distributions, or leveraging homology with model organisms.

2.4.5 Differential Activation Analysis

We fit a log-normal distribution to the differences in iModulon activities between biological replicates for each iModulon. For a single comparison, we computed the absolute value of the difference in the mean iModulon activity and compared it against the iModulon's log-normal distribution to determine a p-value. We performed this comparison (two-tailed) for a given pair of conditions across all iModulons at once and designated significance as FDR < 0.01.

2.4.6 Data and Code Availability

All data generated or analyzed during this study are included in the published article (and its Supplementary Information Files). The original data set is from Nicolas, et al. [50] (GEO accession number GSE27219; Supplementary Data from http://genome.jouy.inra.fr/basysbio/bsubtranscriptome/). Interactive online dashboards for all iModulons and all data are available at https://iModulonDB.org under the data set name "*B. subtilis* Microarray". Code for the analysis pipeline used here is available on GitHub (https://github.com/SBRG/precise-db).

2.5 Acknowledgements

We thank Dr. Joe Pogliano, Dr. Saugat Poudel, and Dr. Eammon Riley for helpful discussions and biological insights. We also thank Juan Tibocha-Bonilla and Dr. Karsten Zengler for their interest in a follow-up study which explored the results in **Sections 2.2.2.1** and **2.2.2.6** in a metabolic and expression model [108] This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was funded by the Novo Nordisk Foundation Center for Biosustainability (Grant Number NNF10CC1016517).

Author Contributions: K.R. analyzed data and drafted the paper; A.V.S. designed research; A.V.S. and B.O.P. provided mentorship and guidance throughout. All participated in writing the paper.

Chapter 2, in part, is a reprint of the material as appears in: **Rychel K**, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome. *Nat Commun.* 2020 Dec 11; 11(1):6338. The dissertation author was the primary author.

Chapter 3. iModulonDB: a Knowledgebase of Microbial Transcriptional Regulation Derived from Machine Learning

Independent component analysis (ICA) of bacterial transcriptomes has emerged as a powerful tool for obtaining coregulated, independently-modulated gene sets (iModulons), inferring their activities across a range of conditions, and enabling their association to known genetic regulators. By grouping and analyzing genes based on observations from big data alone, iModulons can provide a novel perspective into how the composition of the transcriptome adapts to environmental conditions. Here, we present iModulonDB (iModulonDB.org), a knowledgebase of microbial transcriptional regulation computed from high-quality transcriptomic datasets using ICA. Users select an organism from the home page and then search or browse the curated iModulons that make up its transcriptome. Each iModulon and gene has its own interactive dashboard, featuring plots and tables with clickable, hoverable, and downloadable features. This site enhances research by presenting scientists of all backgrounds with co-expressed gene sets and their activity levels, which lead to improved understanding of regulator-gene relationships, discovery of transcription factors, and the elucidation of unexpected relationships between conditions and genetic regulatory activity. The original release of iModulonDB covered three organisms (*E. coli*, *S. aureus*, and *B. subtilis*) with 204 iModulons, and it was later expanded to cover eleven organisms and 1,717 iModulons.

3.1 Background

The transcriptional regulatory network (TRN) governs gene expression in response to environmental stimuli, which is of fundamental interest in biology. The TRN functions by employing condition-responsive regulators, such as transcription factors (TFs), to regulate the transcription of genes. Understanding these regulators and their effects in bacteria informs many important applications, including bioproduction [49] and antibiotic resistance [109]. There are several organism-specific databases of gene-regulator relationships [11], [12], [110], but knowledge of regulator binding alone is insufficient to explain the complex responses reflected in transcriptomic datasets [15], [111]. Thus, there is a need for data-driven approaches to TRN elucidation, which can detect the most important transcriptional signals in a gene expression dataset, identify their major gene constituents, and quantify their condition-dependent activity levels.

With the growing availability of bacterial transcriptomes, machine learning is emerging as a powerful tool for TRN elucidation. The falling price of RNA sequencing has led to a rapid growth in online transcriptomic databases [112], [113], creating a strong need for the development of analytical tools that can harness its scale to transform raw data into biologically meaningful information [114]. For transcriptomic data, this knowledge comes in the form of 1) identifying which regulons are active in each condition probed in the dataset, 2) generating hypotheses about gene function and regulation, and 3) revealing novel relationships and patterns in bacterial lifestyles. In comparison, traditional methods such as chromatin immunoprecipitation (ChIP) assays [115], can be time-consuming and expensive, making them cumbersome for high-throughput discovery or hypothesis generation. They also do not yield the condition-specific strength of binding, which can be inferred by machine learning. Another strength of data-driven approaches is that they can be applied to any organism, regardless of prior information. Ultimately, a comprehensive, quantitative TRN would be the result of this pursuit.

Independent component analysis (ICA) addresses the goals described above. It is a blind source separation algorithm that identifies statistically independent signals underlying a dataset, and decomposes the original matrix into source (or module, **M**) and activity (**A**) matrices [22], [23]. The module matrix **M** defines relationships between genes and the identified signals, while the activity matrix **A** describes the intensity of each signal in each sample. A comparison of 42 TRN inference methods demonstrated that ICA was the best at recovering known gene modules [17]. Additional studies found that ICA-derived gene modules were robust across datasets [38]. It has been used for a variety of organisms in the past, especially yeast and human cancer cells [33]–[37].

We recently applied ICA to high quality transcriptomic datasets for three species of bacteria: *Escherichia coli* [19], *Staphylococcus aureus* [116], and *Bacillus subtilis* [117]. In our work with *E. coli*, we termed the independent signals “iModulons”, for *i*ndependently *modul*ated gene sets. We used the same codebase (www.github.com/SBRG/precise-db) to generate all three transcriptome decompositions. We then assigned categories, functions, and regulators to each iModulon. The regulator assignments were based on existing knowledge of transcription factor binding sites [11], [12], [110], [116], or, in some cases, a boolean combination of several regulons. By comparing our decompositions to known regulons described in other databases, we were able to identify potentially mislabeled or previously unknown TF-gene associations and verify them with ChIP-exo binding profiles. For example, we discovered five new binding sites for MetJ in *E. coli* [19]. The observed iModulon activity levels also provided additional evidence for newly discovered regulatory roles or point to new hypotheses. For instance, we observed that the iron chelator pulcherrimin was active in stationary phase signaling in *B. subtilis* (**Section 2.2.2.4**) [117]. This observation was recently externally validated [75]. More broadly, each decomposition explained 65-80% of the variance in the transcriptomic data sets, indicating that the functions and regulators identified captured most of the transcriptional functions of the TRNs under the conditions where the data was obtained.

Several other studies have utilized iModulons to obtain valuable results. Our original *E. coli* study led to the discovery of a regulon putatively controlled by the uncharacterized TF *ydhC* [19], which has since been characterized and renamed [118]. iModulons were also used to characterize the function of the TF OxyR [40], and to quantify the stress responses to heterologous gene expression [41]. Additionally, iModulons captured the transcriptional effect of genomic alterations in adaptive laboratory evolution to naphthoquinone-based aerobic respiration [39].

In previous publications, curated iModulon dashboards were presented as static supplemental PDFs. While these are useful for disseminating basic information on specific iModulons, they suffered from several problems: 1) labeling of genes and inclusion of tables were limited by page space; 2) the inability to interact via hovering and clicking slowed analysis; 3) information was static and could not be updated; 4) access to the underlying data was limited and required coding experience; and 5) dissemination required sharing large PDFs instead of simple URL links. Given the potential for iModulons to enhance a wide range of research, we sought to address these issues with an online knowledgebase.

Here, we present the iModulon database (iModulonDB; iModulonDB.org), an interactive web tool for accessing high quality transcriptomic datasets and their curated iModulons. Users begin by selecting their organism and dataset of interest, and then may either browse the iModulons identified through curated tables or search for the genes and regulators of interest to them. Each gene and iModulon has an interactive analytics dashboard featuring hoverable, clickable, and downloadable tables and graphs. iModulonDB presents the relationships between genes and iModulons, the inferred activities of regulators across diverse conditions, and the concordance between our data-driven gene modules and literature-defined regulons, such as those available on RegulonDB [11]. Compared to the previous PDF dashboards, the new website enables global usage of this information and provides much more depth. iModulonDB meets the emerging need for an online, data-driven TRN resource; it obtains co-regulated gene sets based only on transcriptomic observations, which will help a broad audience of microbiologists and systems biologists to investigate genes or iModulons of interest.

3.2 Materials and Methods

3.2.1 Data Generation and Acquisition

E. coli PRECISE-278 [19] and *Staph*PRECISE [116] were generated using RNA sequencing (RNAseq), as described in their respective publications. The *Bacillus subtilis* dataset [117] is a well-known microarray dataset originally published by Nicolas, et al. [50], and is already featured as the expression compendium on the popular database *Subti*Wiki [12]. Despite using older, established data, the *Bacillus* decomposition still provides novel insights (**Chapter 2**), which demonstrates the efficacy of ICA. All raw data is available on GEO or SRA.

3.2.2 Quality Control and Preprocessing

Prior to running ICA, we ensure that transcriptomic data passes stringent quality control as described in each dataset's original publication [19], [116], [117]. For RNAseq data, genes shorter than 100 nucleotides or with under 10 fragments per million-mapped reads are removed. We then compute transcripts per million (TPM) using DESeq2 [28]. The final expression compendium is log-transformed $\log_2(\text{TPM} + 1)$ before analysis – this is the form of the data available for download on iModulonDB. For both RNAseq and microarray data, biological replicates with $R^2 < 0.9$ between final expression values are removed to reduce technical noise. Before computing iModulons, we choose a reference condition (such as exponential growth on M9 minimal media) and subtract its expression from all other samples; this results in activity levels for each iModulon relative to a known baseline.

3.2.3 Computing Robust iModulons

We perform ICA as described in the original publications [19], [116], [117] using the Scikit-learn [24] implementation of the FastICA algorithm [119]. To ensure robustness (since ICA is a stochastic gradient search algorithm), we perform ICA multiple times with random seeds for each dataset and cluster their **M** matrices using the Scikit-learn implementation of the DBSCAN algorithm [120]. We keep independent components that appear in more than 50% of the ICA runs. The code used to compute robust independent components is publicly available (github.com/SBRG/precise-db). Note that previous publications used **S** to refer to the **M** matrix, but it has since been renamed to avoid confusion with the stoichiometric matrix **S** [4].

ICA produces a set of independent components, each of which contains a weight for each gene in the expression dataset. Most gene weights are near zero for a given independent component, so the genes with large positive or negative weights are considered to be in the iModulon. To transform an independent component into an iModulon, we iteratively remove the highest absolute weighted genes from an iModulon until the D'Agostino K^2 statistic of normality [106] of the remaining distribution falls below an organism-specific cutoff, and take all removed genes to be iModulon members. See the original publications (and **Chapter 2**) for additional details [19], [116], [117].

3.3 Results

3.3.1 A Web-based Analytics Platform for Data-Driven TRNs

We developed iModulonDB (iModulonDB.org) to enhance the field of microbial genetic regulation by presenting TRNs based on observed signals in transcriptomic datasets. This site provides biologists with the ability to easily navigate large datasets and quickly find gene modules through a search tool or by browsing curated annotations. The inferred activity levels of each iModulon, which are readily available on the site, provide valuable, novel insights into cellular function. We hope that iModulonDB will become an important part of the database ecosystem, providing a machine learning-derived perspective that also links to other databases for synergetic TRN characterization. **Figure 3.1** demonstrates the relationship between the content of an iModulon page (bottom left) and the original decomposition (top), as well as the types of insights that may be gleaned from this analysis (bottom right).

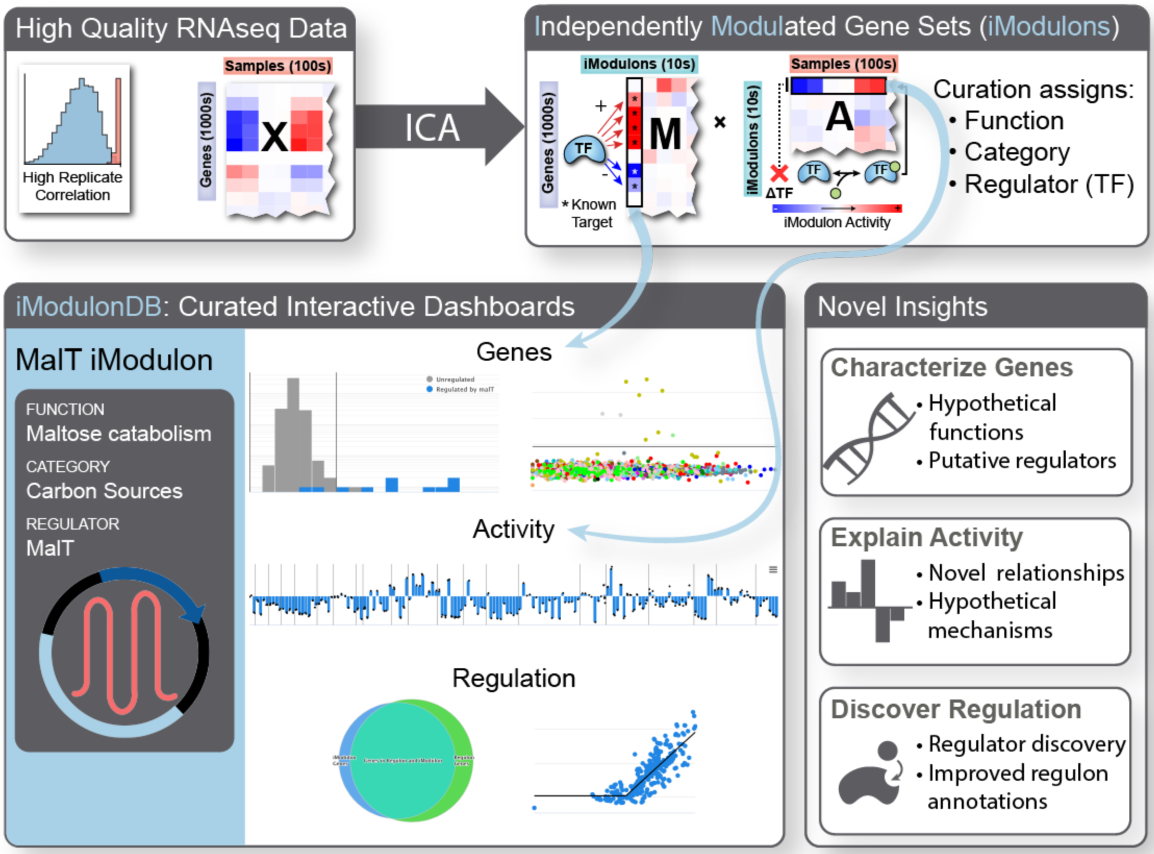


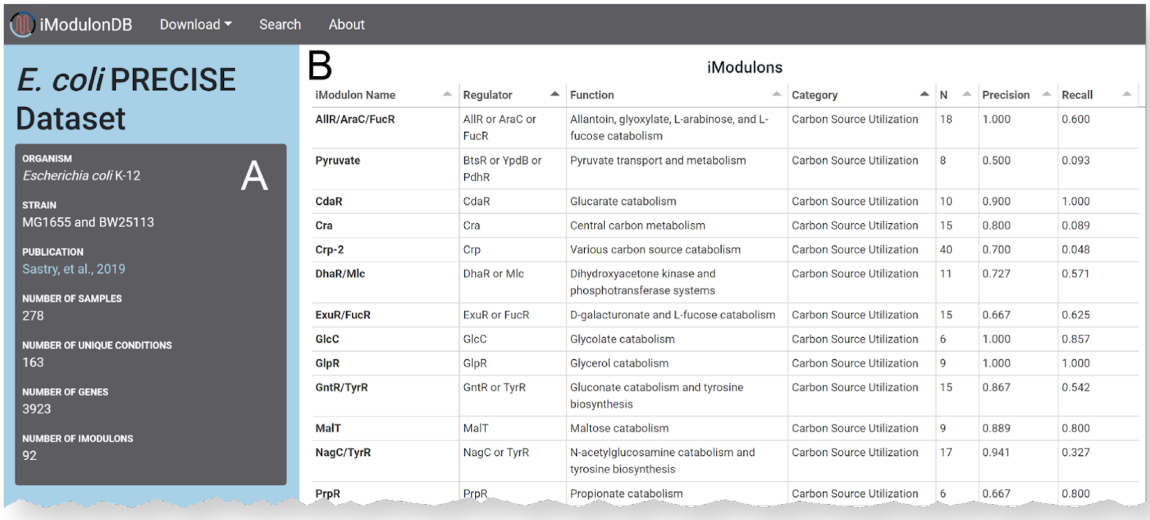
Figure 3.1: General outline of the iModulonDB pipeline. Analysis begins with transcriptomic data **X** (upper left), which is quality controlled for high biological replicate correlation. ICA is used to obtain iModulons, which have gene weights for each gene in the matrix **M** and activity levels for each condition in the matrix **A** (upper right). **M** weights are analogous to the strength of transcription factor binding upstream of genes and **A** activities are analogous to condition-dependent transcription factor activities, which may depend on processes such as ligand binding. Note that **X** ≈ **M** * **A**. The iModulons are curated by assigning functions, categories, and regulators, as shown in the example MalT iModulon dashboard. A representative screenshot of the interactive graphs is shown (middle; larger version with labels included in **Figure 3.3**), covering genes and activities as well as the concordance between this gene set and its curated regulator. The three rows of the dashboard result in the three categories of novel insights described (bottom).

iModulonDB originally contained the three datasets listed in **Table 3.1**. It covered 204 iModulons, 180 of which were characterized. The *E. coli* dataset is the largest; it includes 278 expression profiles with various gene knock-outs and evolved strains. This leads to its high dimensionality and lowest explained variance. The presence of genomic alterations also results in the most genomic iModulons [121], reducing the fraction of iModulons with known regulators. The *S. aureus* dataset is comparatively smaller, and is explained very well by its iModulons. Nearly all iModulons could be characterized, but the dearth of existing TRN annotations for this organism makes it more difficult to align certain iModulons to regulators. For *B. subtilis*, both mRNAs and non-coding RNAs were included in the decomposition. This leads to a high number of "genes" and 17 uncharacterized (noisy) iModulons. However, 95% of characterized iModulons were associated with a known set of regulators. Overall, the decompositions successfully produced iModulons that align well with our existing knowledge of transcriptional regulation in these species.

Table 3.1: Representative statistics of the three datasets originally in iModulonDB. 'Genes' and 'Samples' refer to the dataset size. Note that in *B. subtilis,* non-coding RNAs were included in the tiling array, which leads to the high gene count. 'Conditions' are unique experimental conditions, not counting biological replicate samples. 'Dimensionality' is the number of orthogonal principal components needed to explain 99% of the variance in the data. The 'Characterized iModulons' row lists values and percentages of iModulons that can be categorized and named (excluded iModulons may result from noise or contain genes for which little information is available). 'iModulons with Regulators' are the subset of characterized iModulons that are mapped to regulators (excluded iModulons may be effects of genomic alterations or biological enrichments with no known regulators). 'Explained Variance' is the fraction of variance explained by reconstructing the original matrix using only iModulon member genes and activities. SBRG: Systems Biology Research Group.

	E. coli [19]	*S. aureus* [116]	*B. subtilis* [117]	Total
Dataset Description	RNA sequencing at SBRG	RNA sequencing at SBRG	Tiling Microarray, Nicolas, *et al.* [50]	
Genes	3923	2820	5875	12618
Samples	278	108	265	651
Conditions	163	54	104	321
Dimensionality	200	73	95	368
Total iModulons	92	29	83	204
Characterized iModulons	86 (93.5%)	28 (96.6%)	66 (79.5%)	180 (88.2%)
iModulons with Regulators	58 (63.0%)	19 (65.5%)	63 (75.9%)	140 (68.6%)
Explained Variance	68%	76%	72%	

3.3.2 Dataset Pages Show Lists of iModulons, Constituting a Data-Driven TRN



E. coli PRECISE Dataset

ORGANISM
Escherichia coli K-12

STRAIN
MG1655 and BW25113

PUBLICATION
Sastry, et al., 2019

NUMBER OF SAMPLES
278

NUMBER OF UNIQUE CONDITIONS
163

NUMBER OF GENES
3923

NUMBER OF IMODULONS
92

iModulons

iModulon Name	Regulator	Function	Category	N	Precision	Recall
AllR/AraC/FucR	AllR or AraC or FucR	Allantoin, glyoxylate, L-arabinose, and L-fucose catabolism	Carbon Source Utilization	18	1.000	0.600
Pyruvate	BtsR or YpdB or PdhR	Pyruvate transport and metabolism	Carbon Source Utilization	8	0.500	0.093
CdaR	CdaR	Glucarate catabolism	Carbon Source Utilization	10	0.900	1.000
Cra	Cra	Central carbon metabolism	Carbon Source Utilization	15	0.800	0.089
Crp-2	Crp	Various carbon source catabolism	Carbon Source Utilization	40	0.700	0.048
DhaR/Mlc	DhaR or Mlc	Dihydroxyacetone kinase and phosphotransferase systems	Carbon Source Utilization	11	0.727	0.571
ExuR/FucR	ExuR or FucR	D-galacturonate and L-fucose catabolism	Carbon Source Utilization	15	0.667	0.625
GlcC	GlcC	Glycolate catabolism	Carbon Source Utilization	6	1.000	0.857
GlpR	GlpR	Glycerol catabolism	Carbon Source Utilization	9	1.000	1.000
GntR/TyrR	GntR or TyrR	Gluconate catabolism and tyrosine biosynthesis	Carbon Source Utilization	15	0.867	0.542
MalT	MalT	Maltose catabolism	Carbon Source Utilization	9	0.889	0.800
NagC/TyrR	NagC or TyrR	N-acetylglucosamine catabolism and tyrosine biosynthesis	Carbon Source Utilization	17	0.941	0.327
PrpR	PrpR	Propionate catabolism	Carbon Source Utilization	6	0.667	0.800

Figure 3.2: Representative screenshot of the dataset page for *E. coli* PRECISE-278. **(A)** Details of the dataset, including a link to the relevant publication. **(B)** List of iModulons for this dataset, along with some curated details and statistics. Click a row to access the appropriate iModulon page. Screenshot includes the first 13 iModulons on the table, which all happen to be in the category ‘Carbon Source Utilization’.

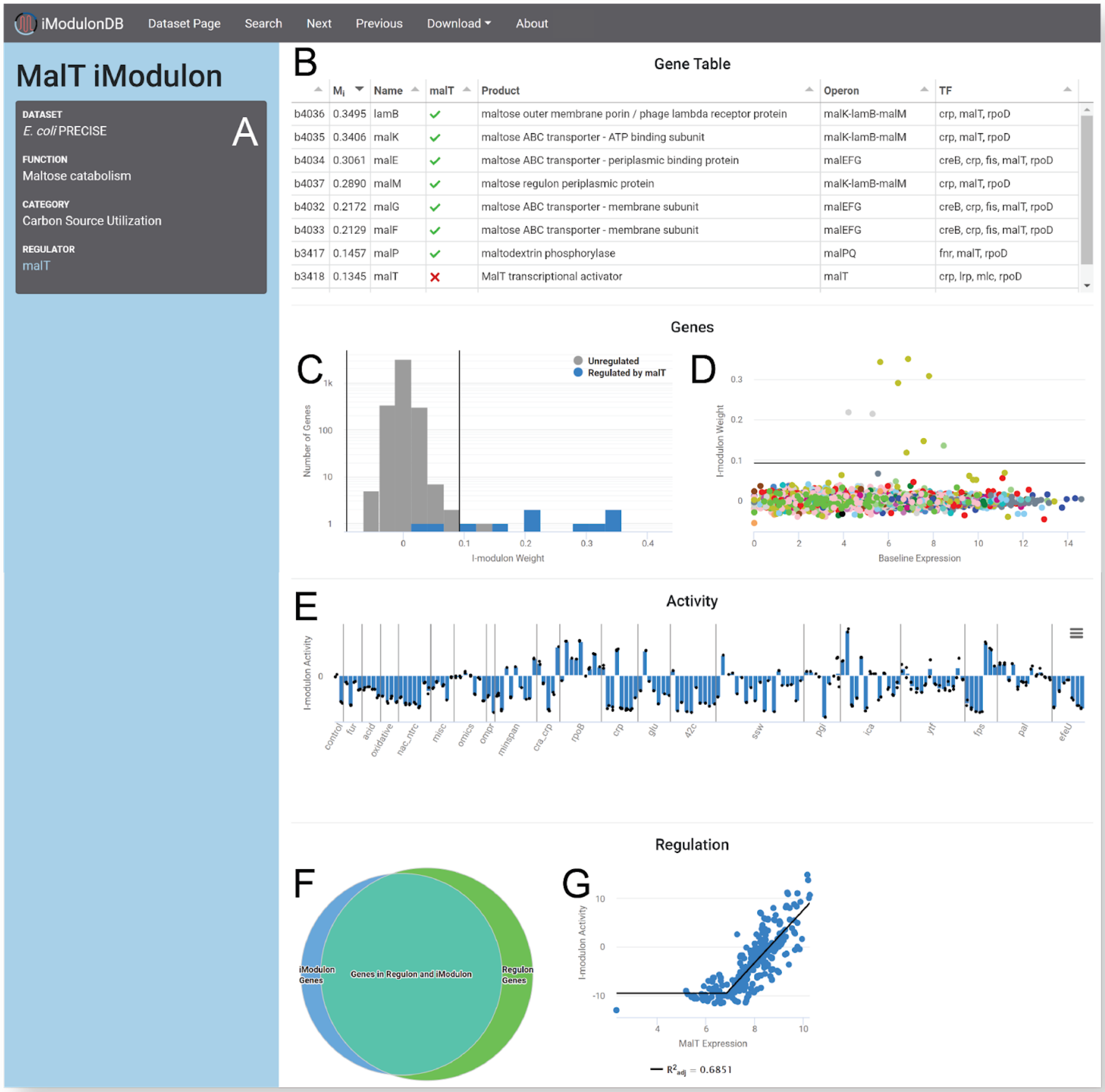
Users begin by selecting one of our decompositions from the home page. This brings them to a dataset page (**Figure 3.2**). The left sidebar (**Figure 3.2A**) lists basic features of the dataset, including organism details, dimensions, and a link to the relevant publication. The page also contains a curated table of all the iModulons identified (**Figure 3.2B**). iModulons are given a name, regulator, function, and category. When multiple regulators capture the iModulon better than a single one, they may be combined using “or” (/) or “and” (+) set operators. For example, the ExuR/FucR iModulon contains genes regulated by either ExuR or FucR, while the GutM+SrlR iModulon contains only genes regulated by both GutM and SrlR (see the original publications [19], [116], [117] for more details). If multiple iModulons are linked to the same regulator, which indicates subsets of a large regulon, then a hyphen and numeral is added to the iModulon name (e.g. Fur-1). Some iModulons are enriched for a function or genomic alteration instead of a regulator, so they are named accordingly. Three statistics are also included in the table: 1) N, the number of genes; 2) precision, the fraction of iModulon genes regulated by the enriched regulator(s); and 3) recall, the fraction of regulated genes in the iModulon. Users may sort the table by any of its columns, or click on a row to learn about the iModulon.

Some common iModulon categories are: carbon source utilization, amino acid and nucleotide biosynthesis, metal homeostasis, stress response, and lifestyle (such as biofilm production). There are also iModulons categorized as 'uncharacterized', which may indicate undiscovered genetic or regulatory relationships [121].

3.3.3 iModulon Pages Present Information-Dense Interactive Analytic Dashboards

After selecting an iModulon, users are taken to its iModulon page (**Figure 3.3**). A gray box on this page (**Figure 3.3A**) lists the curated features: function, category, and regulator. If available, the regulator entry will contain links to other databases (RegulonDB in *E. coli* [11] or *Subti*Wiki in *B. subtilis* [12]). If a regulator is assigned, the precision and recall of the enrichment is shown.

The gene table (**Figure 3.3B**) contains information about all gene members of the iModulon. By default, the gene table is sorted by the associated gene weight from the **M** matrix (although users may click any column to sort by a different feature). Most columns contain information from external databases (EcoCyc in *E. coli* [122], AureoWiki in *S. aureus* [123], and *Subti*Wiki in *B. subtilis* [12]), such as gene product descriptions, operons, and associated regulators. If the iModulon has been assigned regulator(s), then the regulator names will appear as columns in the table, with green checks if the gene is known to be associated with that regulator and red X's if not. This table is valuable for understanding the iModulon, as browsing the gene function list helps to understand the function of the iModulon as a whole. The boolean transcription factor columns are also a tool for discovery, since the genes marked with red 'X's may be controlled by the regulator despite the lack of a known association. Rows are also links to gene pages.



MalT iModulon

DATASET
E. coli PRECISE

FUNCTION
Maltose catabolism

CATEGORY
Carbon Source Utilization

REGULATOR
malT

Gene Table

	M_i	Name	malT	Product	Operon	TF
b4036	0.3495	lamB	✓	maltose outer membrane porin / phage lambda receptor protein	malK-lamB-malM	crp, malT, rpoD
b4035	0.3406	malK	✓	maltose ABC transporter - ATP binding subunit	malK-lamB-malM	crp, malT, rpoD
b4034	0.3061	malE	✓	maltose ABC transporter - periplasmic binding protein	malEFG	creB, crp, fis, malT, rpoD
b4037	0.2890	malM	✓	maltose regulon periplasmic protein	malK-lamB-malM	crp, malT, rpoD
b4032	0.2172	malG	✓	maltose ABC transporter - membrane subunit	malEFG	creB, crp, fis, malT, rpoD
b4033	0.2129	malF	✓	maltose ABC transporter - membrane subunit	malEFG	creB, crp, fis, malT, rpoD
b3417	0.1457	malP	✓	maltodextrin phosphorylase	malPQ	fnr, malT, rpoD
b3418	0.1345	malT	✗	MalT transcriptional activator	malT	crp, lrp, mlc, rpoD

Figure 3.3: Representative screenshot of the iModulon page for MalT from *E. coli* PRECISE-278. **(A)** Details, including a link to its regulator's page on RegulonDB. **(B)** Gene table, listing the members of this iModulon and their annotations, plus whether or not they are annotated as being regulated by the iModulon's regulator, MalT. **(C)** Histogram showing the distribution of gene weights. Vertical bars represent the thresholds between member genes and non-member genes. All member genes are positively weighted in this example. **(D)** Scatter plot of gene weights versus their expression in the baseline condition, color coded by COG. The horizontal line represents the weight threshold. **(E)** Activity bar graph for this iModulon. Vertical lines separate projects within the dataset, bars represent mean activity of conditions, and individual black dots represent samples. **(F-G)** This row only exists if the iModulon has a curated regulator. **(F)** Venn diagram comparing this set of iModulon genes (left) to the set of genes annotated as being regulated by MalT (right). **(G)** Scatter plot comparing the expression of MalT to the iModulon activity for all conditions. A broken line is used to indicate that there is a minimum expression of MalT before the correlation is observed.

The gene histogram (**Figure 3.3C**) shows the distribution of weights associating this iModulon with all genes. It has a typical shape with a high number of non-member genes near zero and a small number of member genes outside of the vertical threshold lines. All bars can be hovered over to see a list or number of genes. If the iModulon is assigned multiple regulators, the histogram is color-coded by regulon – regulated genes with weights near zero may have other regulators such that those genes are not part of this independent signal in the transcriptome. Each element in the legend can be clicked to show or hide regulated genes.

The gene scatter plot (**Figure 3.3D**) places the weights on the Y axis and the gene start site on the X axis. If genes share similar X axis positions, they are near each other and may be part of the same operon. Each point is color-coded by cluster of orthologous groups (COG; computed via eggNOG [124], [125]), hovering over a point displays additional information, and clicking takes you to the corresponding gene page. This plot allows you to visualize COG categories and view gene annotations for both member and non-member genes.

The activity bar graph (**Figure 3.3E**) displays the activity level of the iModulon (with respect to a reference condition) across all expression profiles. iModulon activities are a measure of relative transcription factor activity, which can be a very valuable resource that is difficult to obtain using other methods. Bars represent conditions and dots represent individual samples within that condition. Hovering over a bar displays some relevant metadata about culture conditions, and clicking it may take you to a paper with more details if it exists (for example, the *E. coli* dataset contains expression profiles from multiple projects, and each project has its own paper). Clicking the wrench next to 'Activity' brings up a menu to select which metadata details are included in the hover box, and allows the selection of a metadata category to color the bars with. The graph can also be zoomed in and scrolled horizontally, or viewed in full screen mode.

If the iModulon has a matched regulator, then a 'Regulation' row will appear at the bottom of the dashboard (**Figure 3.3F-G**), which quantifies the concordance between our transcriptomics-derived groupings (iModulons) and the gene groupings (regulons) available in literature and on other databases

[11], [12]. It contains a venn diagram comparing the set of iModulon genes against the regulon (**Figure 3.3F**), which can be hovered over to quickly see agreements and discrepancies between iModulons and regulon annotations. If the regulator is also encoded by a gene in the expression compendium, then there will be a scatter plot or series of scatter plots showing the iModulon activity and regulator gene expression across the conditions (**Figure 3.3G**). Each point can be hovered over to see the name of the sample it represents. For some regulators, such as sigma factors and self-regulating TFs, a high correlation may be observed in this plot. When post-transcriptional modifications such as phosphorylation and ligand binding affect the iModulon activity, we see low correlations between the expression of the regulator and the activity of the genes it regulates (See **Section 2.2.2.8**). Regulators are not necessarily part of their own iModulons. The computation of these correlations is described in the original dataset publications [19], [116], [117]. In some cases, the correlation is best captured using a broken line (as in **Figure 3.3G**) to signify that a minimum expression level of the regulator must be reached before a correlation is observed.

3.3.4 Gene Pages Connect Users to iModulons of Interest

Gene pages enable users to quickly find iModulons relevant to their research. Similar to the iModulon and dataset pages, the left sidebar on this page will list the gene’s identifier, gene name, gene product, operon, COG category, and known regulator(s), as well as a link to a relevant database (EcoCyc in *E. coli* [122], Aureowiki in *S. aureus* [123], and *Subti*Wiki in *B. subtilis* [12]). The dashboard contains two elements: 1) a table of iModulons and 2) an expression profile. The expression profile shows $\log_2$ TPM expression values across each dataset.

The iModulon table (included in **Figure 3.4B**) lists the iModulons in order of weight, with the strongest iModulon associations at the top of the list. If the gene is in an iModulon, a green check will appear next to it (red X otherwise). The table also includes additional details for each iModulon. Here, a user can easily find iModulons containing genes of interest, and navigate to the iModulon page to learn more. If a gene is not in any iModulon, the gene is not a strong part of any independent signals in the dataset.

3.3.5 Other Major Features: About, Search, and Download

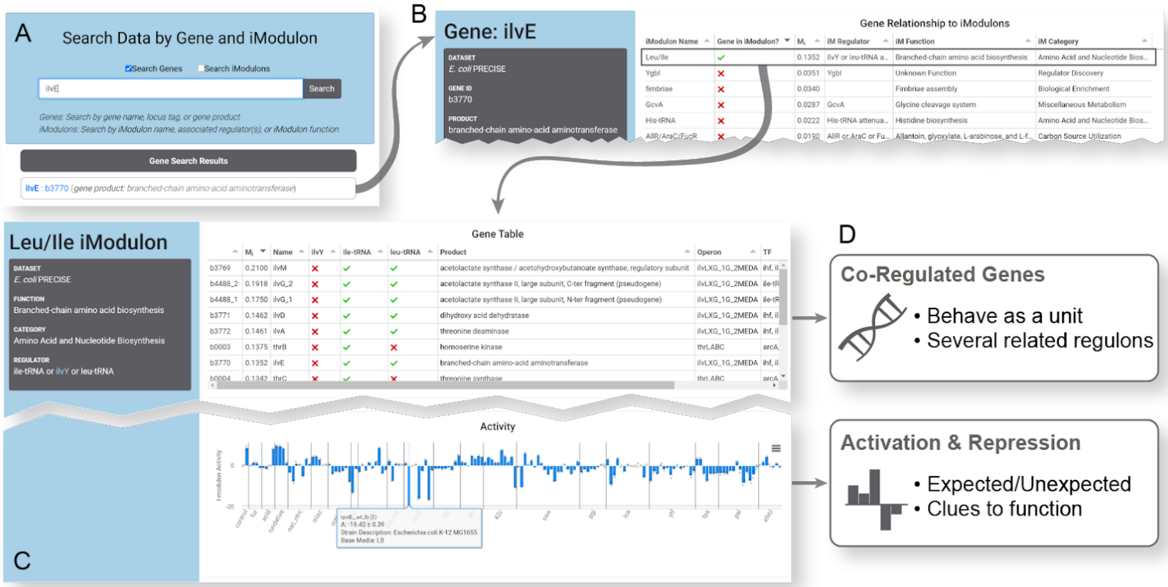
As many researchers are not familiar with iModulons, we have included a thorough “About” page with a YouTube video on ICA, an illustrated walkthrough of our pipeline, details on how to read our dashboards, and links to all relevant publications. We also include an email address (iModulonDB@ucsd.edu) where users can send feedback or request that we work with their dataset.

After choosing a dataset, users may click the “Search” link in the toolbar from any page to access our search functionality. This tool allows the users to search through genes and/or iModulons. For genes, results matching queried gene names, gene IDs, and gene products will be displayed. Similarly for iModulons, matching iModulon names, associated regulators, and iModulon functions will generate search results. Results are separated into iModulon and gene sections.

Each of the main three pages (dataset, iModulon, and gene) has a download menu in the toolbar. From there, all data is available for download in bulk or parts. The bulk data includes the original log-tpm data, the two ICA-generated matrices, the gene annotations and literature TRN used, the metadata on all samples, and the curated iModulon table. For individual iModulons, the gene weights and activity levels are available along with the gene table as it appears in the dashboard. Likewise, gene pages support the download of all iModulon associations to that gene, its expression levels for each sample, and the iModulon table as it appears in the dashboard. We encourage data download and custom analysis.

3.3.6 Case Study: How to Use iModulonDB to Enhance Research

This section includes an example of a researcher who could gain valuable information from iModulonDB, and how they might do so. **Figure 3.4** illustrates the process.



Search Data by Gene and iModulon

☑Search Genes ☐Search iModulons

ilvE Search

Genes: Search by gene name, locus tag, or gene product
iModulons: Search by iModulon name, associated regulator(s), or iModulon function

Gene Search Results

ilvE : b3770 (*gene product: branched-chain amino-acid aminotransferase*)

Gene: ilvE

DATASET
E. coli PRECISE

GENE ID
b3770

PRODUCT
branched-chain amino-acid aminotransferase

Gene Relationship to iModulons

iModulon Name	Gene in iModulon?	M_i	iM Regulator	iM Function	iM Category
Leu/Ile	✓	0.1352	IlvY or leu-tRNA a...	Branched-chain amino acid biosynthesis	Amino Acid and Nucleotide Bios...
YgbI	✗	0.0351	YgbI	Unknown Function	Regulator Discovery
fimbriae	✗	0.0340		Fimbriae assembly	Biological Enrichment
GcvA	✗	0.0287	GcvA	Glycine cleavage system	Miscellaneous Metabolism
His-tRNA	✗	0.0222	His-tRNA attenua...	Histidine biosynthesis	Amino Acid and Nucleotide Bios...
AllR/AraC/FucR	✗	0.0190	AllR or AraC or Fu...	Allantoin, glyoxylate, L-arabinose, and L-f...	Carbon Source Utilization

Leu/Ile iModulon

DATASET
E. coli PRECISE

FUNCTION
Branched-chain amino acid biosynthesis

CATEGORY
Amino Acid and Nucleotide Biosynthesis

REGULATOR
ile-tRNA or ilvY or leu-tRNA

Gene Table

	M_i	Name	ilvY	ile-tRNA	leu-tRNA	Product	Operon	TF
b3769	0.2100	ilvM	✗	✓	✓	acetolactate synthase / acetohydroxybutanoate synthase, regulatory subunit	ilvLXG_1G_2MEDA	ihf, il
b4488_2	0.1918	ilvG_2	✗	✓	✓	acetolactate synthase II, large subunit, C-ter fragment (pseudogene)	ilvLXG_1G_2MEDA	ile-tR
b4488_1	0.1750	ilvG_1	✗	✓	✓	acetolactate synthase II, large subunit, N-ter fragment (pseudogene)	ilvLXG_1G_2MEDA	ile-tR
b3771	0.1462	ilvD	✗	✓	✓	dihydroxy acid dehydratase	ilvLXG_1G_2MEDA	ihf, il
b3772	0.1461	ilvA	✗	✓	✓	threonine deaminase	ilvLXG_1G_2MEDA	ihf, il
b0003	0.1375	thrB	✗	✓	✗	homoserine kinase	thrLABC	arcA,
b3770	0.1352	ilvE	✗	✓	✓	branched-chain amino-acid aminotransferase	ilvLXG_1G_2MEDA	ihf, il
b0004	0.1342	thrC	✗	✓	✗	threonine synthase	thrLABC	arcA,

rpoB__wt_lb (2)
A: -19.40 ± 0.39
Strain Description: Escherichia coli K-12 MG1655
Base Media: LB

Co-Regulated Genes
- Behave as a unit
- Several related regulons

Activation & Repression
- Expected/Unexpected
- Clues to function

Figure 3.4: Flowchart of an example analysis for the gene *ilvE*. **(A)** The user searches for their gene of interest, *ilvE*. This returns the gene as a result, which can be clicked to access its gene page (B). **(B)** The iModulon table on the gene page shows that this gene is the member of the Leu/Ile iModulon. Clicking that row in the table brings the user to the iModulon page (C). **(C)** Two features of the iModulon page are shown: the gene table and activity bar graph (see **Figure 3.3** for details). The activity bar graph includes an example tooltip, which appears when the bar is hovered over with a mouse. In this case, the bar with the tooltip is a repressing condition labeled with LB media, which is expected to repress the function of this iModulon. **(D)** The insights gained are summarized in the boxes: the co-regulated genes are enumerated, and the activating or repressing conditions in our dataset can be determined.

Researcher X has chosen a gene of interest. They are studying the branched chain amino acid (BCAA) synthesis enzyme encoded by *ilvE* in *E. coli* as a potential drug target. They want to know which conditions activate *ilvE* expression, and what other genes are co-expressed with this gene. Researcher X can navigate to the *E. coli* dataset on iModulonDB, and then choose “Search” in the toolbar and enter “ilvE” (**Figure 3.4A**). *ilvE* will appear as a gene page result, which links to the *ilvE* gene page (**Figure 3.4B**). From there, they can see some basic information about the gene, a link to its EcoCyc page [122], and its expression across our compendium. They will also find the iModulon table, which shows that it is a member of the “Leu/Ile” iModulon. As expected, the function of the iModulon is BCAA biosynthesis. Clicking the “Leu/Ile” row in the iModulon table will take them to the corresponding iModulon page (**Figure 3.4C**) where they can see all of the genes (including *ilvE*) in this independent signal of the dataset, as well as its activity across all conditions. The iModulon is annotated as being regulated by ile-tRNA, ilvY, or leu-tRNA. IlvY is a transcription factor with a page on RegulonDB [11], so users can follow that link to learn more about its regulon.

Along with linking to relevant pages on other databases, iModulonDB provides Researcher X with some novel information (**Figure 3.4D**). The first would be the particular gene grouping of this iModulon – it covers the synthesis of all three BCAAs plus threonine as a single unit of the transcriptome, instead of three separate regulons or several operons as they may find on other databases. This grouping is observed from transcriptomic data, and likely results from a combination of genetic factors that respond to separate but metabolically related metabolites. It is useful to understand the cell as manipulating all of these genes together as a unit, which may affect how the researcher implements their drug targeting. The other major insight would be gained by probing the activity profile; for example, hovering over the tallest bars shows that this iModulon is activated under iron starvation and reactive oxygen species (ROS) stress and hovering over the most negative bars indicates deactivation under rich media conditions or under osmotic stress from NaCl. Searching the literature for explanations would reveal that ROS and iron starvation damage the iron-sulfur clusters needed for member enzymes [126] and create BCAA-limiting conditions (See **Section 4.2.8.2**), while the amino acids in rich media turn off these genes through their well-studied mechanisms [127]. Direct or indirect repression of these genes by osmotic stress has not been studied and may be worth investigating. All of this information would help Researcher X determine whether *ilvE* and the function encoded by the Leu/Ile iModulon are good potential drug targets.

Not all iModulons need to be found through gene pages. Users may be more interested in a general process or specific regulator, which can also be searched for on the search page. Alternatively, users can start by selecting an iModulon from the dataset page. They may stumble upon a surprising gene grouping or unexpected condition/iModulon activation pair, which could lead to valuable new hypotheses worth testing. This knowledgebase would also be very valuable for someone studying a completely uncharacterized gene, as its presence in an iModulon gives clues to its function. Another way to use this knowledgebase is simply to follow along with the existing publications [19], [116], [117]. For

E. coli and *B. subtilis*, expected and unexpected activating conditions for each iModulon were listed in supplementary tables, which may serve as a starting place for hypothesis generation.

3.3.7 Design and Implementation

iModulonDB is implemented and deployed using a simple web application stack and combination of user interface technologies. The server side is entirely hosted by GitHub Pages (pages.github.com) with HTTPS enforced. It relies on local computations performed in Python 3.7 with Jupyter notebooks (jupyter.org). The client side is implemented using a combination of HTML, CSS, and JavaScript. Bootstrap 4.5 (getbootstrap.com) is used to manage page layout and ensure mobile compatibility. Interactive tables were made using Tabulator (tabulator.info), and plots are implemented in HighCharts using a non-commercial license (highcharts.com). Other JavaScript packages used include: jQuery (jquery.com), Popper.js (popper.js.org), URLSearchParams (developer.mozilla.org), and PapaParse (papaparse.com).

3.4 Discussion

iModulonDB is a data-driven bacterial TRN knowledgebase that meets the need for unsupervised, observation-based gene groupings and inferred genetic regulator activities. Its interactive graphical user interface with search and download functionality facilitates usage by scientists with computational and non-computational backgrounds alike. The platform provides a novel perspective on bacterial genomes based on big data observations, and also connects to (and largely agrees with) databases that report established gene groupings based on past literature. It originally included data for three organisms based on three publications and contains 204 iModulons, and was later expanded to accommodate new datasets and organisms (See **Section 3.6**). This work demonstrates the potential for iModulonDB to improve our understanding of bacterial TRNs for a wide variety of organisms.

3.5 Data Availability

iModulonDB is freely available online at https://iModulonDB.org and can be accessed with a JavaScript-enabled browser. The download links in the toolbars enable download of all data and facilitate custom analysis.

3.6 Continued Growth of iModulonDB

iModulonDB expanded rapidly after its original publication, as shown in **Figure 3.5**. To facilitate rapid growth, a pipeline was set up [128] to scrape all available RNAseq data from the SRA database [18], [113]. Thus, data from past research could be reused and reanalyzed through the iModulon framework, which revealed important insights at this new scale. In some cases, additional samples were generated in-house in order to supplement the available datasets and probe specific questions.

As of the completion of this dissertation, iModulonDB contains 11 organisms, which span the phylogenetic tree of life (**Figure 3.4A**) [129]. In addition to the original three organisms, iModulonDB covers *Acinetobacter baumannii* [130], *Mycobacterium tuberculosis* [131], *Pseudomonas aeruginosa* [132], [133], *Pseudomonas putida* [134], *Salmonella enterica* [135], *Sulfolobus acidocaldarius* [136], *Streptococcus pyogenes* [137], and *Vibrio natriegens* [138]. These diverse organisms include several major human pathogens (*A. baumannii, M. tuberculosis, P. aeruginosa, S. aureus, S. enterica,* and *S. pyogenes*), important bioindustrial species (*P. putida, E. coli,* and *B. subtilis*), and even an archeal extremophile (*S. acidocaldarius*).

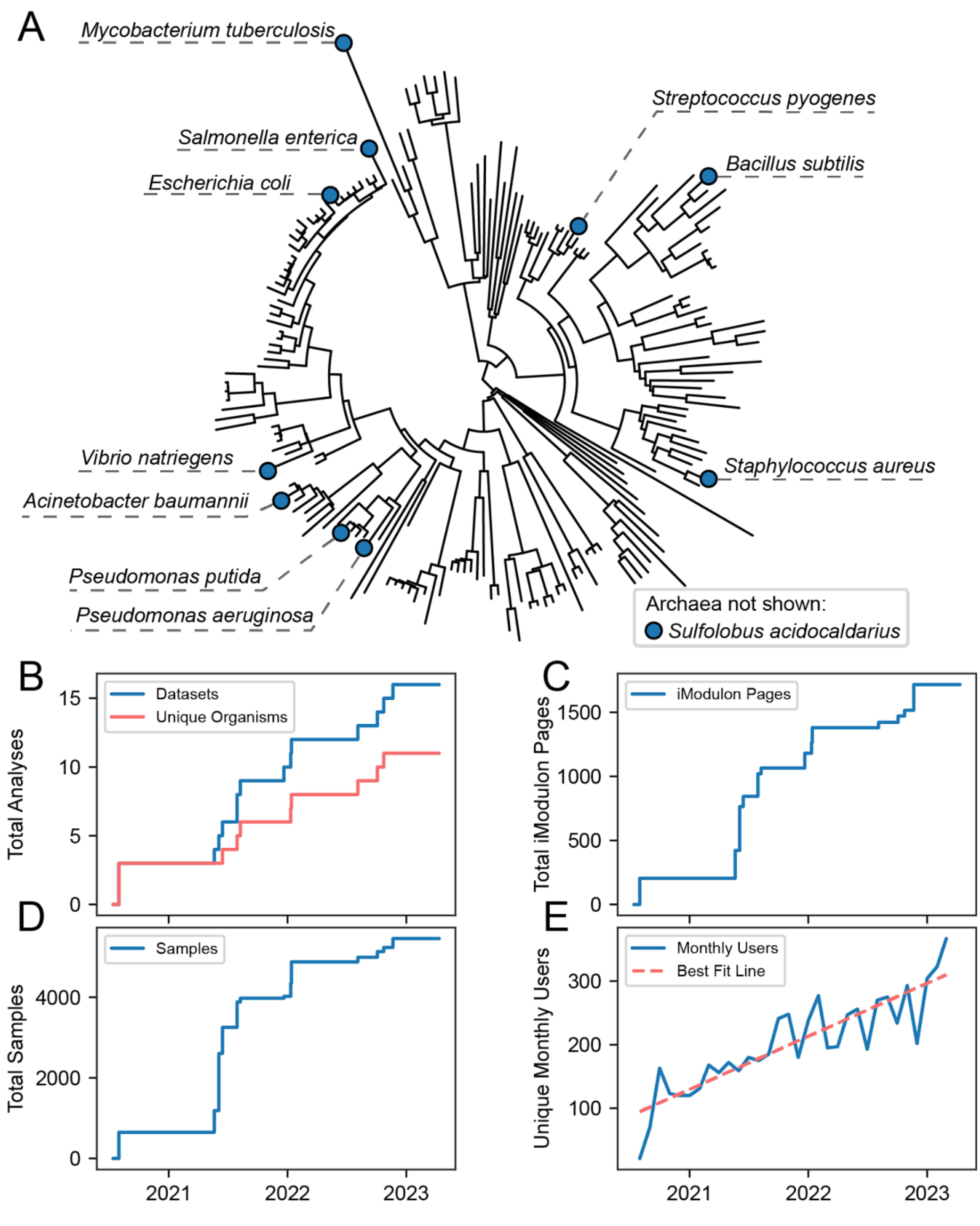


Figure 3.5: iModulonDB rapidly expanded in its first two years after publication. **(A)** Phylogenetic tree representing each of the species currently in iModulonDB, which span the bacterial web of life and includes an archaeal species as well. Adapted from [139] using data from [129]. **(B-E)** Growth of **(B)** datasets (16), organisms (11), **(C)** total iModulon pages (1717), **(D)** total samples analyzed (5456), and **(E)** unique monthly users (367). The best fit line in (E) shows an increase of approximately 7 users per month, representing steady growth and adoption.

The papers associated with these datasets include a wealth of fascinating insights. Here, we highlight a few examples. iModulons provide nuance to the boundaries of biosynthetic gene clusters (BGCs) by revealing coregulated genes, addressing an important shortcoming of existing BGC algorithms with applications for new molecule discovery [132]. Stationary phase transcriptional allocation was revealed in *P. putida*, and comparisons between stationary phase expression and expression under various carbon sources informs growth preferences with potential industrial impact [134]. iModulon structures for different *Salmonella* strains exhibited different gene sets for virulence-associated regulators, providing insights into the genetic and transcriptional basis for virulence [135]. Finally, iModulons revealed the expression changes underlying competency and facilitated discovery of new competency-related genes, an important process to enhance genetic engineering [138]. With all of these structures readily available on iModulonDB, any researcher can easily discover new details previously hidden within these complex datasets and perform follow-up investigations.

In addition to the new organisms, iModulonDB also added a significantly expanded *E. coli* dataset, PRECISE-1K, and an even larger ‘K-12’ dataset that includes all available high quality RNAseq data for *E. coli* K-12 [140]. **Chapters 4** and **5** describe detailed analyses of subsets of samples from PRECISE-1K. The updated iModulons maintain most of the iModulons from PRECISE-278, but increase the total number from 93 to 201. The similarity in structure highlights the robustness of ICA, while the new and refined signals demonstrate that it does improve with additional data.

The growth in various metrics of database size and adoption are shown in **Figure 3.5B-E**. iModulonDB usage has grown steadily since its inception, and was accessed by 367 unique users in March of 2023 (**Figure 3.5E**). The sustained and increasing usage of iModulonDB suggests that it has become a valuable resource to the research community. iModulonDB will continue to be expanded, with several new organisms being analyzed as of this writing.

3.7 Acknowledgements

The authors would like to thank Dr. Akanksha Rajput and David Bour for their input on website infrastructure and code. We would also like to thank Dr. Hyungyu Lim, Dr. Sonal Choudhary, and Reo Yoo for testing the site and providing valuable suggestions. As iModulonDB has continued to grow, we would like to thank all authors who contributed their hard work to building up this resource. We would also like to thank Dr. Jason Hyun for providing the phylogenetic tree which we adapted for this dissertation (**Figure 3.5A**) from his unpublished work.

Author Contributions: K.R. led site development and drafted the paper. K.R. and K.D. developed the website. A.V.S., P.V.P. and B.O.P. provided mentorship and guidance in design and implementation. A.V.S., S.P. and K.R. provided, analyzed, and reformatted data for the site. All participated in writing the paper.

Chapter 3, in part, is a reprint of the material as appears in: **Rychel K,** Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res.* 2021 Jan 8; 49(D1):D112-D120.

Chapter 4. Laboratory Evolution, Transcriptomics, and Modeling Reveal Mechanisms of Paraquat Tolerance

Relationships between the genome, transcriptome, and metabolome underlie all evolved phenotypes. However, it has proved difficult to elucidate these relationships because of the high number of variables measured. A recently developed data analytic method for characterizing the transcriptome can simplify interpretation by grouping genes into independently modulated sets (iModulons). Here, we demonstrate how iModulons reveal deep understanding of the effects of causal mutations and metabolic rewiring. We use adaptive laboratory evolution to generate *E. coli* strains that tolerate high levels of the redox cycling compound paraquat, which produces reactive oxygen species (ROS). We combine resequencing, iModulons, and metabolic models to elucidate six interacting stress tolerance mechanisms: 1) modification of transport, 2) activation of ROS stress responses, 3) use of ROS-sensitive iron regulation, 4) motility, 5) broad transcriptional reallocation toward growth, and 6) metabolic rewiring to decrease NADH production. This work thus reveals the genome-scale systems biology of ROS tolerance.

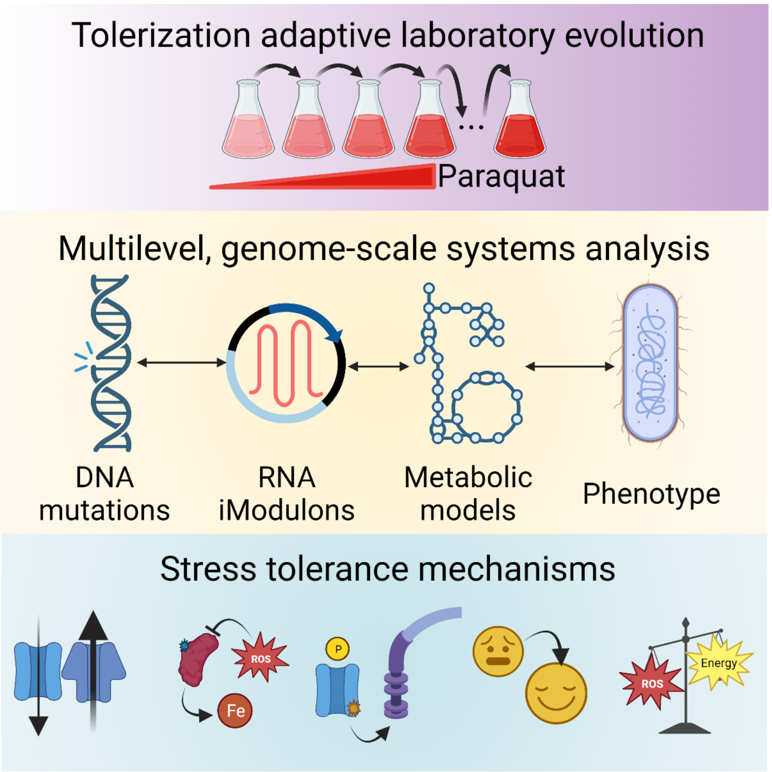

Tolerization adaptive laboratory evolution
Paraquat
Multilevel, genome-scale systems analysis
DNA mutations
RNA iModulons
Metabolic models
Phenotype
Stress tolerance mechanisms
ROS
Fe
P
ROS
Energy

Figure 4.1: Graphical abstract of **Chapter 4**. Adaptive laboratory evolution (ALE) is applied to generate paraquat-tolerized strains, and then a multilevel genome-scale systems analysis is applied, which includes iModulon analysis, analysis of mutations, metabolic modeling, and phenotypic characterization experiments. The result of the analysis of these strains is a set of novel stress tolerance mechanisms which protect the evolved strains from paraquat and oxidative stress.

4.1 Background

Omics technologies have enabled global understanding of cellular states at each level of the central dogma of biology. In particular, the falling cost of nucleotide sequencing has led to a dramatic increase in available genomic and transcriptomic datasets, allowing researchers to probe nucleotide changes in DNA and condition-dependent expression changes in RNA at unprecedented scale [141]. With genome-scale metabolic models, we can also gain a global perspective on metabolic fluxes, and how they change based on genetic or expression perturbations [5], [126], [142]. Each tool on its own has been successful in gaining novel biological insights, but an even deeper understanding can be achieved if they are made interoperable. Many approaches to integrate multiple omics data types are being developed [143], but the high number of variables and employment of complex “black-box” computational tools presents a problem for elucidating a clear, genome-scale understanding of biological systems across multiple levels of genomic, transcriptional, metabolic, and phenotypic changes.

Adaptive laboratory evolution (ALE) is an experimental procedure in which a microbial starting strain is grown in a selected condition for many generations, propagating when flasks reach a targeted density during repeated batch growth. This allows selection to enrich for mutant strains with improved fitness under the chosen condition [144]. A tolerization ALE uses this procedure with increasing stressor concentrations, pushing cells to amplify stress tolerance mechanisms [145], thereby generating unique strains which are stress tolerance specialists. ALE strains are an excellent starting point for developing multi-omic approaches because they have a well-defined phenotype which arises from an average of only ~22 mutations (according to a database of such mutations, ALEdb [146]). ALE mutations are highly informative for improving gene annotations, identifying fundamental biological principles and tradeoffs, designing bioproduction strains, and understanding antimicrobial resistance [144], [147]. However, it is difficult to interpret effects of mutations on regulators and enzymes without adding characterization from the transcriptome and metabolome.

The transcriptional regulatory network (TRN) employs transcription factors (TFs) which sense features of the cellular state and regulate the expression of genes in response. As transcriptomic data has been generated in rapidly growing numbers and deposited into online databases, it has become increasingly important to develop scalable methods which enable their interpretation. However, the typical method for transcriptional analysis, differentially expressed gene (DEG) analysis, is cumbersome for complex transcriptomic adjustments due to the high number of DEGs, and it does not easily capture the large-scale structure of the TRN. We seek to integrate signals from the TRN with mutations in the genome via biologically meaningful relationships, which is difficult if we do not first effectively decrease the number of transcriptomic variables.

A recently developed approach addresses this challenge by using independent component analysis (ICA) of large compendia of transcriptomic data to group genes into independently modulated sets (iModulons). The expression level of each group (iModulon activity) is computed in each sample, allowing systematic, large-scale analysis of the transcriptomic effect of adaptation to a new growth condition. Each iModulon is manually curated with predicted regulators and functions, bridging between the quantitative TRN and existing literature. iModulon activity levels can be used to infer the activity of their underlying regulators, and thus enable quantitative interrogation of the cell's sensory systems. This approach has provided valuable insights into the TRNs of *Escherichia coli* [19], [140] and several other organisms [116], [131]–[136], [148]. iModulon analysis is supported by a developed codebase and online knowledgebase (iModulonDB.org) [21], [128], which are publicly available. iModulons have already shown promise for analyzing transcriptional reallocation in tandem with mutations, which revealed important examples of the interplay between the genome and the transcriptome [39], [40], [149]–[151], but more work needs to be done to explain larger fractions of transcriptomic variance by systematically characterizing iModulon changes.

Downstream of the genome and gene expression, the state of the metabolic network is fundamental in determining cellular phenotypes. We have developed genome-scale metabolic and expression (ME) models, which compute optimal steady-state fluxes for all known reactions in a cell given mathematical constraints and an objective function [5], [152]. These models can be constrained with growth rates, uptake and secretion rates from metabolomic data, and transcriptomic data [149], [150]. Recent work has also incorporated the effects of biochemical stresses [126], [153], [154], enabling understanding of the cellular response to stress. Since ME models integrate phenotypic, metabolic, and transcriptional or proteomic data, they can be useful for supporting or refuting separate predictions made by analyzing genomic alterations.

The goal of the present Chapter was to gain a genome-scale, multilevel, "white-box" understanding of a particular phenotype by leveraging ALE, genome sequencing, iModulons and ME modeling. Thus, we needed to select a well-defined phenotype of interest. We did so by employing ALE to generate *E. coli* strains which are specialized to tolerate a common herbicide, the redox-cycling compound paraquat (PQ). PQ is a redox-cycling compound, meaning that it can generate large amounts of reactive oxygen species (ROS) by stripping electrons from cellular electron carriers, such as NADH and NADPH, and reducing oxygen; this generates destructive superoxide ROS and regenerates the oxidizing agent to re-initiate the cycle [155]–[157]. The ROS are particularly damaging to iron-containing enzymes and DNA. They decrease activity of important pathways, challenge the integrity of the genome, and inhibit growth [126], [157]–[161].

Though the ROS response of *E. coli* is well understood and ROS are often delivered in the laboratory by PQ [159], some questions remain about how high levels of tolerance can be achieved: (i) In addition to the known proteins, which transporters and enzymes are involved in PQ cycling? (ii) What transcriptional alterations, specifically with respect to stress responses, metal homeostasis, and redox balance, are optimal? (iii) How can cells balance a tradeoff between generating NAD(P)H for energy and decreasing its production to prevent stress generation? Through our unique combination of systems biology techniques, we are able to shed new light on these questions. Their answers are informative for the fundamental biology of stress and metabolism, and for applications in pathology, antimicrobial design, and biomanufacturing.

This work provides a blueprint for combining ALE, mutational analysis, transcriptomics, computational biology, and phenotypic characterizations for stress-tolerant ALE strains, which emphasizes the rich insights provided by iModulon analysis. We begin by characterizing the strains and presenting an overview of the genomic and transcriptional changes. We then show that the effects of large DNA changes and TF mutations are easily quantified in the transcriptome. We also find an unexpected non-TF mutation that regulates motility regulons in our strains. Next, we disentangle the large fraction of the transcriptome which responds to changes in stress and growth phenotypes. Finally, we propose and model a metabolic mechanism for PQ tolerance which involves several interesting mutations and broad transcriptional reallocation. We show that the evolved strains employ a multi-pronged strategy of: (i) modifying membrane transport, (ii) using the SoxS and OxyR regulons to ensure stress readiness, (iii) allowing ROS-sensitive iron-sulfur (Fe-S) clusters to play a larger role in regulation of metal homeostasis, (iv) increasing motility, (v) shifting transcriptional allocation toward growth, and (vi) using fermentation to avert the PQ cycle. Taken together, these results elucidate a detailed, coherent, multilevel understanding of an important cellular phenotype by combining several cutting edge technologies in big data analytics and computational biology.

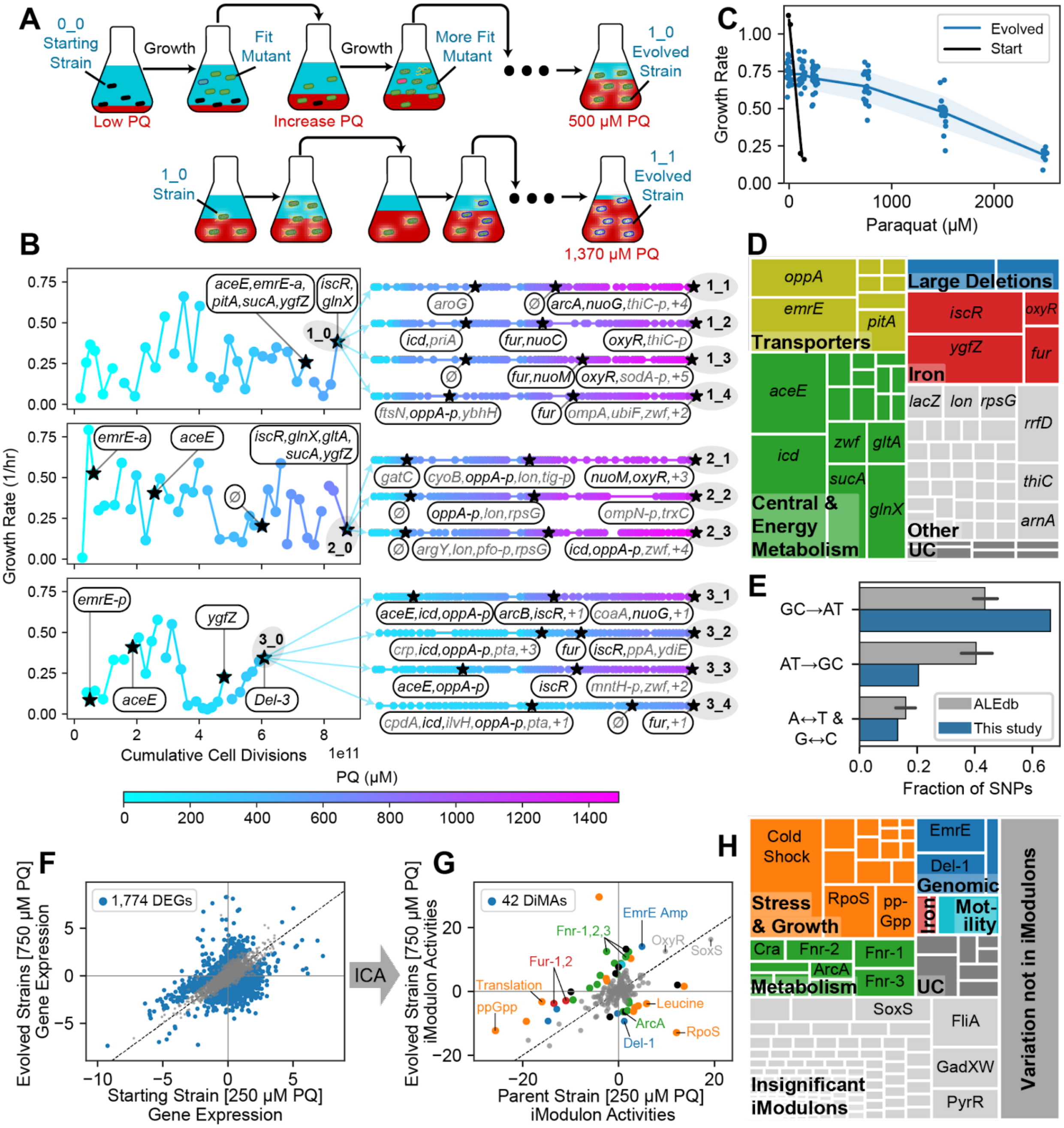
4.2 Results

4.2.1 Laboratory Evolution Increased Tolerated PQ Levels by 1000%

We evolved strains aerobically in minimal media with glucose under increasing PQ stress (**Figure 4.2A**). Our starting strain (0_0) was a derivative of *E. coli* K-12 MG1655 which had been pre-evolved to grow in minimal media with glucose [162]. By using this media-adapted starting strain, the subsequent ALEs were enriched for mutations which improve stress tolerance, since the mutations that promote rapid growth under the culture conditions were already fixed. ALE was performed by steadily increasing PQ concentrations, first in three parallel first generation ALEs (1_0, 2_0, 3_0) and followed by eleven second generation ALEs (1_1, 1_2, …, 2_1, etc.) (**Figure 4.2A-B**). Parallelizing ALE replicates generated diverse strains and allowed for identification of common mutation targets which are more likely to be causal.

After evolution, growth rates for each endpoint under different PQ concentrations were measured (**Figure 4.2C**). The starting strain's growth was severely impaired by low PQ concentrations, with no growth at 250 µM PQ. The evolved strains showed a dramatic increase in the concentration of PQ they can tolerate while still growing; some endpoint strains tolerated 2500 µM. There was a fitness cost to the PQ tolerance, however: the strains no longer grew as well in the absence of PQ as the starting strain. This observation is consistent with the tradeoffs of the PQ tolerization mechanisms.

Figure 4.2: ALE increases PQ tolerance via changes to the genome and transcriptome. **(A)** Tolerization ALE process, showing mutant strains (cells with various appearances) in media with increasing stress concentrations (red). Example replicates are shown: 1_0 in the first generation and 1_1 in the second generation. **(B)** Points represent ALE flasks colored by their PQ concentration. The first generation of ALEs (strains 1_0, 2_0, and 3_0) are shown with each flask's growth rate. "Cumulative cell divisions" are estimated from the growth rate and time elapsed. Stars represent flasks that underwent DNA sequencing, and newly mutated genes are shown. Black colored genes are discussed in detail. **(C)** Growth rate for each strain at each PQ concentration. The starting strain cannot grow at 250 µM PQ, whereas some evolved strains reach up to 2500 µM PQ. Evolved strains grow slower than the starting strain in the absence of PQ. **(D)** Treemap of mutations in all strains, grouped by gene with intergenic mutations assigned to nearest genes. UC: Uncharacterized. **(E)** Fraction of SNP types in this study compared with all public ALE studies on ALEdb (aledb.org; mean ± 95% confidence interval). Each label corresponds to four of the twelve possible substitutions; for instance, "GC→AT" includes "G→A", "G→T", "C→A" and "C→T" substitutions. This study is enriched for mutations which decrease the GC content of the genome. **(F-G)** Comparison between the mean transcriptomes of the parent strain at 250 µM PQ vs. all evolved strains at 250 and 750 µM PQ. **(F)** DEG analysis, showing an intractably large number of DEGs. **(G)** Differential iModulon activity (DiMA) analysis, which compresses the differential transcriptomic changes into 42 DiMAs. DiMAs are colored by their category from panel (H). For more information about each iModulon, explore the PRECISE-1K *E. coli* dataset at iModulonDB.org. **(H)** Treemap of the explained variance of each iModulon in the transcriptome of the evolved strains. The map is first broken into three parts: the colorful region, composed of iModulons that are differentially activated after the evolution and categorized, the light gray region composed of iModulons that do not show a significant trend with evolution, and the dark gray region, representing the error in the iModulon decomposition.


A
0_0 Starting Strain
Growth
Fit Mutant
Growth
More Fit Mutant
1_0 Evolved Strain
Low PQ
Increase PQ
500 μM PQ
1_0 Strain
1_1 Evolved Strain
1,370 μM PQ
C
Growth Rate
Evolved
Start
Paraquat (μM)
B
Growth Rate (1/hr)
aceE,emrE-a, pitA,sucA,ygfZ
iscR, glnX
1_0
1_1
1_2
1_3
1_4
aroG
Ø
arcA,nuoG,thiC-p,+4
icd,priA
fur,nuoC
oxyR,thiC-p
fur,nuoM
oxyR,sodA-p,+5
ftsN,oppA-p,ybhH
fur
ompA,ubiF,zwf,+2
emrE-a
aceE
iscR,glnX,gltA, sucA,ygfZ
2_0
2_1
2_2
2_3
gatC
cyoB,oppA-p,lon,tig-p
nuoM,oxyR,+3
oppA-p,lon,rpsG
ompN-p,trxC
argY,lon,pfo-p,rpsG
icd,oppA-p,zwf,+4
emrE-p
ygfZ
3_0
aceE
Del-3
3_1
3_2
3_3
3_4
aceE,icd,oppA-p
arcB,iscR,+1
coaA,nuoG,+1
crp,icd,oppA-p,pta,+3
fur
iscR,ppA,ydiE
aceE,oppA-p
iscR
mntH-p,zwf,+2
cpdA,icd,ilvH,oppA-p,pta,+1
fur,+1
Cumulative Cell Divisions
1e11
PQ (μM)
D
oppA
emrE
pitA
Transporters
Large Deletions
iscR
oxyR
ygfZ
fur
Iron
aceE
icd
zwf
gltA
sucA
glnX
Central & Energy Metabolism
lacZ
lon
rpsG
rrfD
thiC
arnA
Other
UC
E
GC→AT
AT→GC
A↔T & G↔C
ALEdb
This study
Fraction of SNPs
F
1,774 DEGs
Evolved Strains [750 μM PQ] Gene Expression
Starting Strain [250 μM PQ] Gene Expression
ICA
G
42 DiMAs
EmrE Amp
Fnr-1,2,3
OxyR
SoxS
Fur-1,2
Translation
ppGpp
Leucine
ArcA
RpoS
Del-1
Evolved Strains [750 μM PQ] iModulon Activities
Parent Strain [250 μM PQ] iModulon Activities
H
Cold Shock
RpoS
ppGpp
EmrE
Del-1
Genomic
Iron
Motility
Stress & Growth
Cra
Fnr-2
ArcA
Fnr-1
Fnr-3
Metabolism
UC
SoxS
FliA
GadXW
PyrR
Insignificant iModulons
Variation not in iModulons

4.2.2 Adaptive Mutations Reflect Effects of PQ

Throughout the PQ ALE, a total of 222 mutations were observed, representing 111 unique sequence changes. Each mutation was assigned to its closest gene in the case of intergenic mutations, and 72 total genes were affected. Mutations were then categorized by their likely effects (**Figure 4.2D**). The largest category of mutated genes was central and energy metabolism-related (35%), which reflect the metabolic effects of PQ on redox balance. Transporters were also frequently mutated (16%), likely to prevent influx or promote efflux of PQ or other ROS. Iron and iron-sulfur (Fe-S) clusters are sensitive to oxidative stress [159], so we observed changes to iron regulators and Fe-S cluster synthesis genes (16%). Three large deletions, Del-1, Del-2, and E14 removal, were also notable (5%). Other mutations which were less convergent across endpoint strains (26%) were observed in ribosomal subunits, tRNAs, and *lon* protease, as well as across other parts of the metabolic network.

We performed DNA sequencing on several midpoint strains during the ALEs (**Figure 4.2B**), which provided insight into the most effective growth strategies since mutations tend to fix in the order of fitness benefit [163]. We note that *emrE* and *aceE* are among the first genes to be affected in all three of our first generation strains.

An interesting pattern arose in the observed single nucleotide polymorphisms (SNPs): compared to other ALE projects available on ALEdb [146], they are highly enriched for changes from guanine or cytosine to adenine or thymine (Figure 4.2E**;** Fisher's exact test $p = 9.38*10^{-5}$). This enrichment was consistent with direct damage to DNA by ROS, since guanine is the most easily oxidized nucleotide [158], [164], [165]. Thus, these mutations might not only improve cellular fitness through genomic and transcriptomic changes, but also by physically tolerizing cellular DNA to oxidation.

4.2.3 iModulons Enable Analysis of Complex Transcriptomic Changes

To identify transcriptomic adaptations, we performed RNAseq on the starting strain at 0 and 250 μM PQ, and on each evolved strain at 0, 250, and 750 μM PQ. In a comparison between the stressed samples for the pool of all evolved strains vs. the starting strain, we found 1,774 differentially expressed genes (DEGs) (**Figure 4.2F**), making detailed analysis using traditional transcriptomic methods challenging. Therefore, we applied iModulon analysis to enable interpretation.

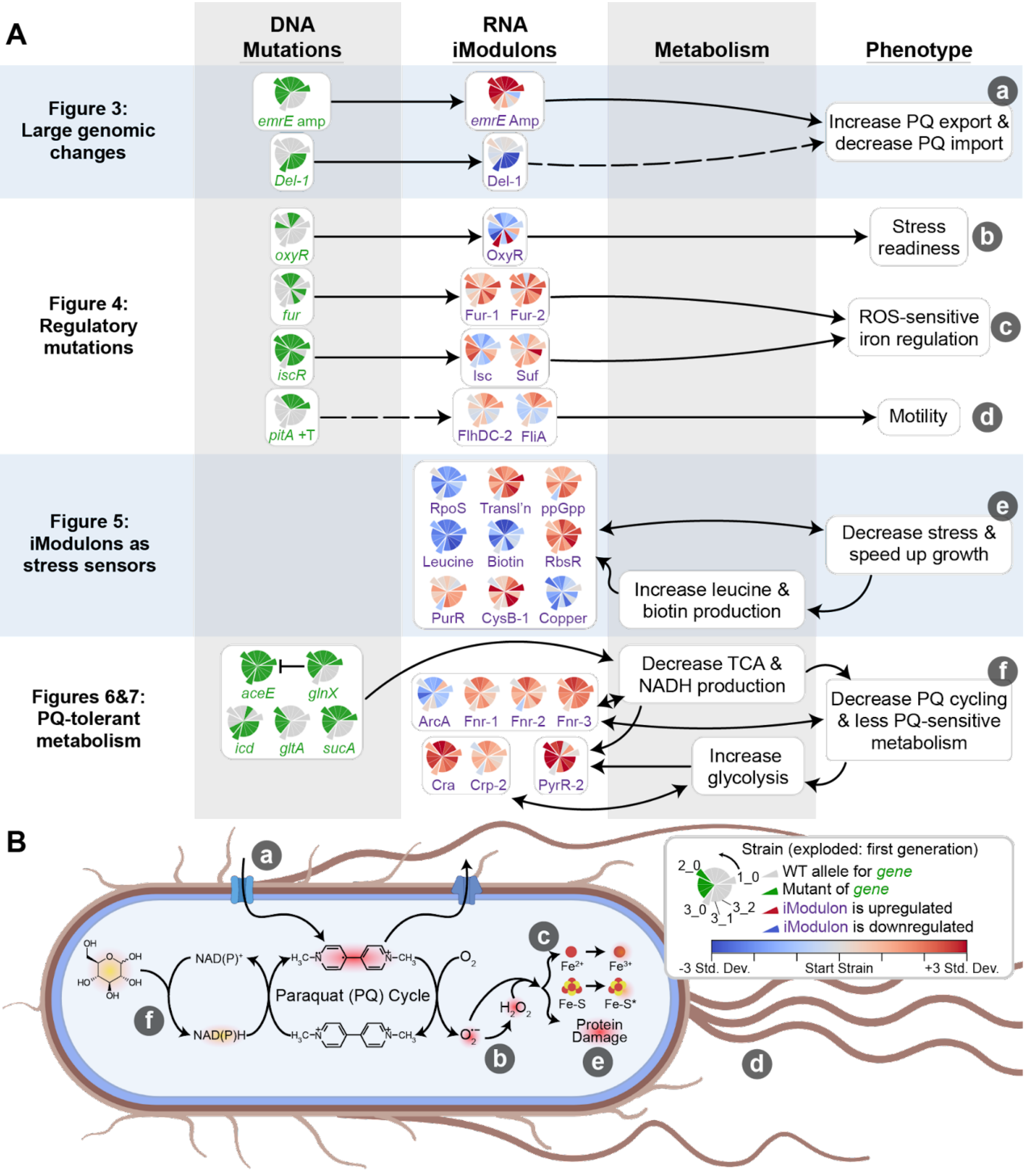
The data was included in a large compendium of *E. coli* RNAseq data generated from a single wet lab protocol (PRECISE-1K [140]). By leveraging over 1,000 samples across diverse conditions, this dataset facilitated machine learning of global transcriptomic patterns. Following our pipeline [128], we performed ICA on the full dataset. The result was a set of 201 iModulons, independently modulated gene sets which have similar expression patterns, along with their activities in each sample. Together, the iModulons constitute a quantitative regulatory structure which maps well to the known TRN, and can be used to reduce the dimensionality of the dataset. The set of PRECISE-1K iModulons was characterized in a separate study [140], and the iModulon structure, including interactive plots, search, and download functionality, is available at iModulonDB.org under *E. coli* PRECISE-1K [21].

iModulons enabled a global characterization of changes in the transcriptome. The evolved strains' gene expression under PQ stress against the starting strain had only 42 statistically significant differential iModulon activities (DiMAs) (**Figure 4.2G**). These 42 iModulons made the analysis of the large-scale changes in the transcriptome tractable, and their observed activity changes could be related to the mutations fixed under ALE. We categorized the DiMAs and assigned mechanistic hypotheses which explain their changes. Explained variance for all categories of significant and insignificant iModulons are shown in **Figure 4.2H**.

4.2.4 A Multilevel Approach on Explaining iModulon Activities Revealed the Effects of Mutations and Phenotypes of Evolved Strains

Modifications to the genome can affect the transcriptome in several ways: large deletions and amplifications can directly alter the expression of genes involved, mutations in TFs can change the expression of their associated regulons, and the transcriptome can adjust due to changes in metabolites or other sensed processes that result from mutations. The latter type of alteration can be complicated by the fact that gene expression also regulates metabolite concentrations and sensed processes. In **Figure 4.3A**, we summarize how each of these types of relationships were observed in the evolved strains. iModulons play a central role in each highlighted mechanism, as evidenced by the full second column in **Figure 4.3A**. Their utility is a key outcome of this work. The combined analysis of genomic and transcriptional changes led us to six key cellular mechanisms of PQ tolerance (**Figure 4.3B**). Together, these mechanisms constitute a summary of the systems biology of PQ-generated ROS stress tolerance.

Figure 4.3: Multilevel approach reveals mechanisms of PQ tolerance. **(A)** Knowledge graph summarizing multilevel relationships between mutations, iModulons, metabolism, and phenotypes. Pie charts appearing in the two left columns indicate prevalence of given changes to the genome and transcriptome (legend in panel B), where wedges indicate strains. The protruding wedges correspond to the first generation of ALEs, with the wedges counterclockwise to them being their second generation descendants. For genes, green indicates the strain has mutations affecting it or its promoter. For iModulons, colors indicate the difference between the iModulon activity in the strain at 750 µM PQ and the starting strain at 250 µM PQ, normalized to the standard deviation of the iModulon activity across all of PRECISE-1K. Dashed lines represent relationships for which there is little existing literature. **(B)** Phenotypic changes target specific processes involved in PQ and ROS stress. Lowercase letters indicate elements from the rightmost column of (A). Entities which glow are reduced, and red indicates stress-related molecules.


A
DNA Mutations
RNA iModulons
Metabolism
Phenotype
Figure 3: Large genomic changes
emrE amp
emrE Amp
Del-1
Del-1
a
Increase PQ export & decrease PQ import
Figure 4: Regulatory mutations
oxyR
OxyR
b
Stress readiness
fur
Fur-1
Fur-2
iscR
Isc
Suf
c
ROS-sensitive iron regulation
pitA +T
FlhDC-2
FliA
d
Motility
Figure 5: iModulons as stress sensors
RpoS
Transl'n
ppGpp
Leucine
Biotin
RbsR
PurR
CysB-1
Copper
e
Decrease stress & speed up growth
Increase leucine & biotin production
Figures 6&7: PQ-tolerant metabolism
aceE
glnX
icd
gltA
sucA
ArcA
Fnr-1
Fnr-2
Fnr-3
Cra
Crp-2
PyrR-2
Decrease TCA & NADH production
f
Decrease PQ cycling & less PQ-sensitive metabolism
Increase glycolysis
B
a
NAD(P)⁺
NAD(P)H
Paraquat (PQ) Cycle
O_2
$O_2^{\cdot-}$
H_2O_2
c
Fe^{2+}
Fe^{3+}
Fe-S
Fe-S*
Protein Damage
b
e
f
d
Strain (exploded: first generation)
2_0
1_0
3_0
3_1
3_2
WT allele for gene
Mutant of gene
iModulon is upregulated
iModulon is downregulated
-3 Std. Dev.
Start Strain
+3 Std. Dev.

4.2.5 Large Amplifications and Deletions in the Genome Affect Membrane Transport

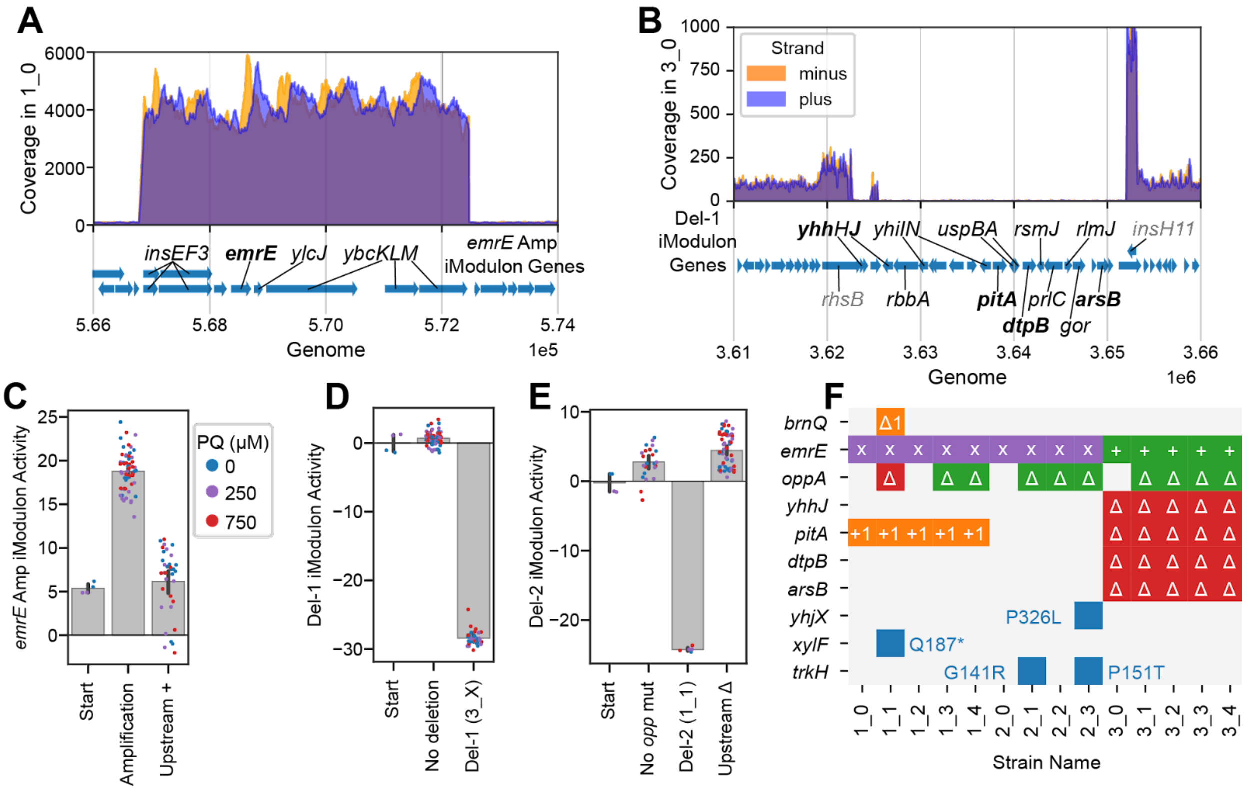


Figure 4.4: Consequences of deletions and amplifications affecting membrane transport are found in both genomes and transcriptomes. **(A)** Genome coverage in strain 1_0, which is representative of strains containing the *emrE* amplification, in the region of the amplification. Genes in the iModulon are labeled. **(B)** Genome coverage of strain 3_0 in the region of Del-1. Del-1 iModulon genes are shown in black, with flanking non-deleted, non-iModulon genes in gray, and transporters in bold. **(C-E)** iModulon activities for selected genomic iModulons. Bars indicate mean ± 95% confidence interval. Individual samples are color-coded by PQ concentration. Upstream + and Δ indicate insertions and deletions, respectively. **(F)** Color-coded table showing all observed mutations related to transporter genes. Purple x: amplification; green: upstream insertion (+) or deletion (Δ); blue: indicated SNP; orange: frameshift mutation within gene; red delta: complete gene deletion. The red area on the right indicates transporters deleted in the major 3_0 deletion.

‘Genomic iModulons’ are transcriptomic modules which capture the effect of large changes to the genome, so they are of primary interest for obtaining genome-to-transcriptome relationships. In the PQ tolerant strains, the major genomic iModulons happen to all be associated with alterations in membrane transport.

The first mutation in each of the first-generation strains affected *emrE*, a multidrug efflux pump which pumps out PQ [166]. In 1_0, 2_0, and their subsequent evolutions, genome coverage was increased approximately 42-fold in the region containing *emrE* (**Figure 4.4A**). This amplification was likely mediated by the flanking DLP12 prophage insertion sequence (IS) genes, specifically the IS3 transposase elements *insEF3* [167]. ICA of the transcriptome recovered the amplified genes as an independent signal in the dataset, which we named the *emrE* Amp iModulon (called ROS TALE Amp-1 in PRECISE-1K [140] on iModulonDB.org [21]). This iModulon showed elevated activity levels in all affected strains regardless of PQ concentration (**Figure 4.4C**). Thus, this case illustrates three levels in our multilevel approach (**Figure 4.3A**): it relates a likely mutational mechanism (transposase-mediated amplification) to a corresponding transcriptomic signal (*emrE* Amp iModulon) and beneficial phenotype (PQ efflux).

In the 3_0 strain and its subsequent evolutions, we do not observe the *emrE* amplification. However, the mutation caller predicted a 9-base pair (bp) insertion 39 bp upstream of *emrE* in these strains, consistent with IS1 insertions that can affect transcription or translation [168]. We do not observe an iModulon signal in the transcriptome of these strains (**Figure 4.4C**). However, we do have evidence that increased expression of *emrE* provides an evolutionary benefit. Therefore, we hypothesize that this mutation would increase translation of EmrE.

The 3_0 strain and its descendants have a large deletion containing 26 genes (**Figure 4.4B**). The deletion may have been mediated by the *insH11* transposase at its 3' end. Similarly to the *emrE* Amp iModulon discussed in the text, the Del-1 iModulon captured the effect of this change in the genome on the composition of the transcriptome. It showed a strong decrease in activity in the strains harboring the deletion (**Figure 4.4D**). The deleted segment contained a variety of genes, making it difficult to deduce its benefit to ROS tolerization. However, we note that it contained four transporter genes: *yhhJ, pitA*, *dtpB,* and *arsB*. Removal of one or several of these transporters may have decreased PQ influx or helped to prevent influx of other oxidized molecules that resulted from oxidative damage.

In addition to the transporters in the Del-1 iModulon, other deleted genes may have been important for the PQ tolerance of the 3_0 strain and its subsequent evolutions (**Figure 4.4B**). These include universal stress response regulators *uspBA*, reductases *gor, arsC,* and *yhiN*, or ribosome-related genes *rbbA, rsmJ,* and *rlmJ*. *yhhJ* and *yhiN* are uncharacterized genes with putative assignments, and these results support their potential role in PQ stress.

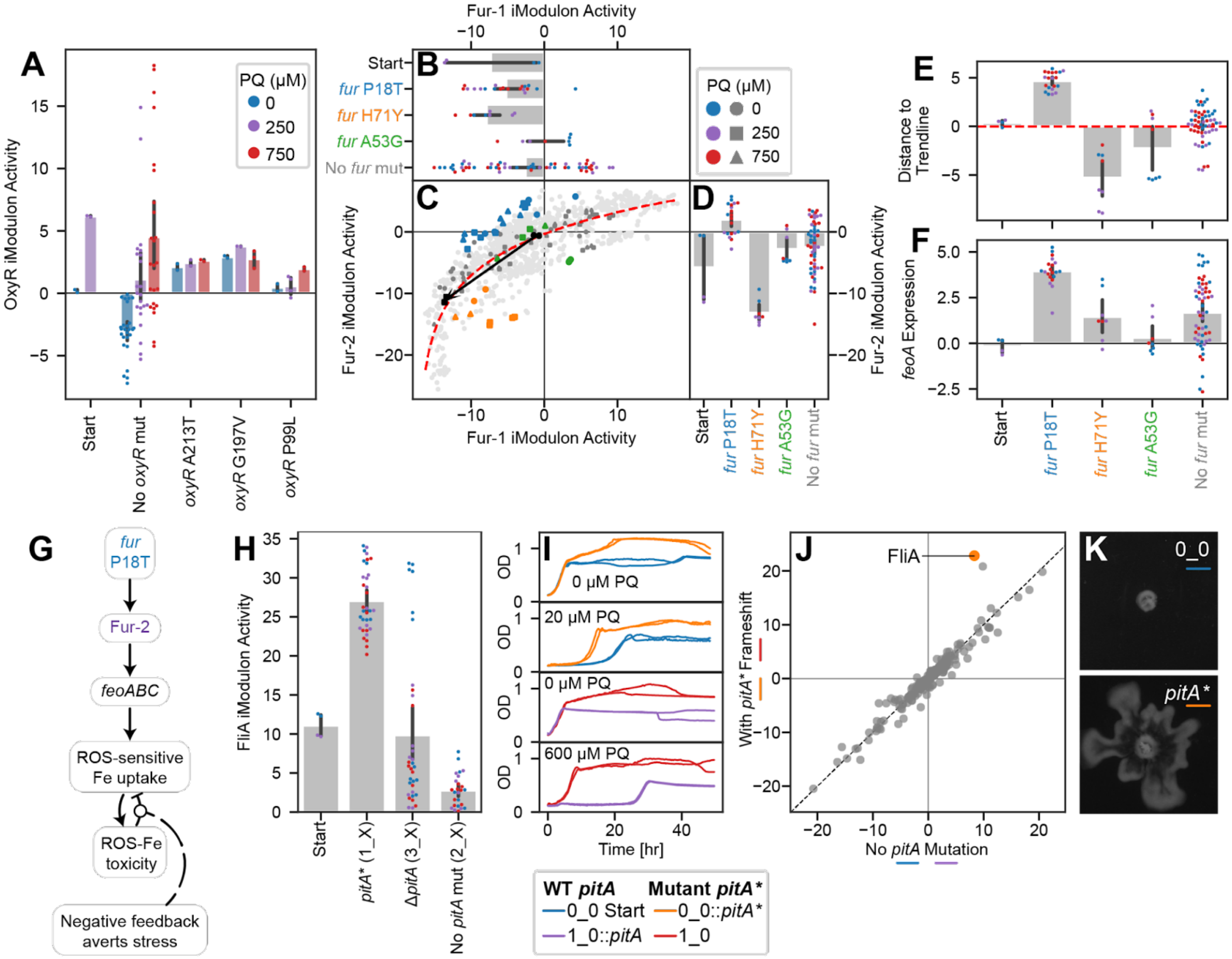
The *oppABCDF* operon was a common target of mutations. Nine of the eleven second-generation strains acquired the same 1,199 bp deletion of the *insH21* IS5 element upstream of it, and one strain, 1_1, deleted the entire operon and its surrounding genes. The deletion was captured by an iModulon (Del-2). The activity of this iModulon shows a downregulation in the deleted strain, and little change between the evolved strains with and without the upstream deletion (**Figure 4.4E**). Since *oppABCDF* is known to be a promiscuous tripeptide transporter that prefers positively charged substrates [169] , it should be considered as a possible route of entry for PQ. The prevalence of the upstream deletion suggests that such a deletion provides improved tolerance, and there is an apparent benefit to a complete deletion of the entire operon. This leads us to predict that the upstream deletion negatively impacts *oppABCDF* translation, as has been suggested in past studies [170], [171].

In addition to the genome-transcriptome-phenotype associations we analyze in depth, mutations on their own can predict putative new functions for their target genes. Therefore, we include all transporters mutated in this study in **Figure 4.4F** so that further research can explore their affinities for PQ and other oxidized compounds, as well as the effects of the observed SNPs.

4.2.6 Mutations in TFs Alter the Regulation of Stress Responses and Iron Homeostasis

'Regulatory iModulons' are iModulons which are statistically enriched with genes from a specific regulon, and their activity level quantifies the activity of the underlying TF. Thus, iModulon analysis reveals the effects of TF mutations in a convenient way.

Figure 4.5: Mutations regulate stress response, iron metabolism, and motility iModulons in novel ways. Bars indicate mean ± 95% confidence interval. **(A)** OxyR iModulon activity is correlated with PQ in starting and evolved strains (Pearson R = 0.47, p = $6.2*10^{-5}$), except for the three strains which mutated oxyR. PQ colors in the legend also apply to panels (B, D, E-F, H). **(B-D)** Scatter plot of Fur-1 and Fur-2 iModulon activities with bar plots sharing axes. Light gray dots indicate other samples from PRECISE-1K. In **(C)**, samples are colored by relevant mutations, and shapes indicate PQ concentrations according to the legends. A black arrow connects the starting strain samples between 0 and 250 µM PQ. In bar plots, point colors indicate PQ concentrations and label colors match with the scatter plots. The red trend line is a logarithmic curve fit to all samples in PRECISE-1K. Samples with the P18T mutation are above the trend line, indicating a preference for Fur-2. **(E)** Distances from each sample in this study to the trend line in (B), more clearly showing the preference for Fur-2 induced by P18T. **(F)** *feoA* expression, which is representative of the *feoABC* operon. Genes are upregulated by the *fur* P18T mutation. **(G)** Knowledge graph linking fur mutation to negative feedback which averts stress. **(H)** FliA iModulon activities by *pitA* mutation, showing an upregulation in the case of the frameshift *pitA**, but not in the case of *pitA* deletion. **(I)** Growth curves for strains with and without the *pitA** mutation as the only difference. The mutation contributes to higher final ODs under no stress, and shorter lag and faster growth under stress. **(J)** DiMA for strains 0_0 and 1_0 with and without the *pitA* frameshift mutation under PQ stress. Points indicate the mean of all relevant samples (individual conditions in duplicate; n=6 per axis). The strains with the mutation significantly activate FliA, one of the motility iModulons. The point near FliA is FlhDC-2, the other major motility iModulon. **(K)** Representative images of swarming in the 0_0 strain with (bottom) and without (top) the *pitA** frameshift. Additional plots: **Figure B.1**; Images for all swarming experiments: **Figure B.2**.


A
OxyR iModulon Activity
PQ (μM)
● 0
● 250
● 750
15
10
5
0
−5
Start
No oxyR mut
oxyR A213T
oxyR G197V
oxyR P99L
B
Fur-1 iModulon Activity
−10
0
10
Start
fur P18T
fur H71Y
fur A53G
No fur mut
PQ (μM)
● ● 0
● ■ 250
● ▲ 750
C
Fur-2 iModulon Activity
0
−10
−20
−10
0
10
Fur-1 iModulon Activity
D
Fur-2 iModulon Activity
Start
fur P18T
fur H71Y
fur A53G
No fur mut
E
Distance to Trendline
5
0
−5
F
feoA Expression
5.0
2.5
0.0
−2.5
Start
fur P18T
fur H71Y
fur A53G
No fur mut
G
fur P18T
Fur-2
feoABC
ROS-sensitive Fe uptake
ROS-Fe toxicity
Negative feedback averts stress
H
FliA iModulon Activity
35
30
25
20
15
10
5
0
Start
pitA* (1_X)
ΔpitA (3_X)
No pitA mut (2_X)
I
OD
1
0
0 μM PQ
20 μM PQ
0 μM PQ
600 μM PQ
0
20
40
Time [hr]
WT pitA
Mutant pitA*
0_0 Start
0_0::pitA*
1_0::pitA
1_0
J
FliA
With pitA* Frameshift
20
10
0
−10
−20
No pitA Mutation
K
0_0
pitA*

4.2.6.1 Fixed OxyR Activity is Achieved by Three Mutations

The OxyR iModulon contains oxidative stress response genes, and its regulator, OxyR, responds to oxidative stress [172]. Thus, we expected its activity level to correlate with PQ level. We found that for most strains, this is the case ($p = 6.2*10^{-5}$). However, we observed three separate *oxyR* mutations which all fix OxyR iModulon activity levels at a level just below that of the stressed starting strain (**Figure 4.5A**), regardless of PQ concentration. We speculate that this level may be ideal because it enables quick detoxification of ROS, while higher levels would be proteomically expensive and/or induce growth-limiting levels of *oxyS* (which is regulated by OxyR and leads to growth arrest [173]). Previous iModulon work in other ALEs found that fixing OxyR in the active conformation provided a fitness benefit [40]. Without the OxyR iModulon to quantify OxyR activity, it would have been much more difficult to define the effect of these mutations.

4.2.6.2 Iron Uptake Adaptation Involves Fur Mutations and ROS-sensitive Iron-Sulfur Clusters

Fur, the ferric uptake regulator, regulates two main iModulons whose activities have a non-linear activity relationship which has been described in detail previously [43] (**Figure 4.5B-D**). Fur-1 mostly contains genes for siderophore synthesis and transport (**Figure B.1A**) which are derepressed under more extreme iron starvation conditions. Fur-2 contains ferrous iron transport genes, as well as siderophore transport and hydrolysis systems, which are derepressed more easily under relatively higher iron concentrations. The activities of the two iModulons form a logarithmic curve (**Figure 4.5C**), which captures the nonlinear effect of Fur on the composition of the transcriptome.

ROS demetallates iron enzymes and oxidizes iron(II) to iron(III) [126], [174]. Thus, PQ would induce higher intracellular iron concentrations that could be sensed by Fur and cause repression of both iModulons (black arrow, **Figure 4.5C**) [175]. This hypothesis is consistent with the starting strain's behavior. After evolution, a decrease in oxidative stress leads to a general upregulation of the Fur-1 and Fur-2 iModulons ($p = 0.031$ and 0.034, respectively).

The evolved strains exhibit a great degree of variation along the Fur curve (**Figure 4.5C**). Since many different factors could perturb iron concentrations for each culture (e.g. local ROS concentrations, trace element mixture variability, enzyme metallation levels, etc.), and Fur is highly sensitive to those concentrations, we believe that this variation is to be expected.

The mutation *fur* P18T was observed in three separate strains (1_2, 1_4, and 3_4). Strains with this mutation tend to be above the trend line in the Fur scatterplot (**Figure 4.5E**), suggesting a higher preference for expressing Fur-2 relative to Fur-1. The strains with this mutation specifically upregulated the *feoABC* genes, which are members of Fur-2 (**Figure 4.5F, B.1A**).This transporter system may be highly beneficial under ROS conditions because it directly couples demetallation of an Fe-S cluster to iron transport, allowing for rapid decreases in iron acquisition when ROS levels are high [176].

Two other mutations were also observed in *fur*. H71Y in 1_3 tends to decrease expression of both iModulons, perhaps by strengthening Fur binding. This would potentially have the benefit of preventing iron toxicity. However, this strategy was not utilized by any other strains and it may have also hampered iron homeostasis in situations where local iron concentrations are low. The other mutation, A53G in 3_2, did not have a detectable effect on the transcriptome.

4.2.6.3 IscR Mutations Modify the Balance of Iron-Sulfur Synthesis Regulons

IscR regulates two separate iron-sulfur (Fe-S) cluster synthesis systems which have iModulons, Isc and Suf [177]. Isc is associated with housekeeping Fe-S synthesis, whereas Suf is robust to iron starvation and ROS stress [178]–[181]. Across our strains, we observed 5 mutations in *iscR*, and each associated with a particular region in a scatter plot of Suf and Isc iModulon activities (**Figure B.1B-D**). Interestingly, most mutations do not strongly upregulate the ROS-tolerant Suf system (**Figure B.1D**), and they either increase or decrease the expression of the Isc system (**Figure B.1B**).

The particular regions in **Figure B.1C** that were selected by the strains are somewhat unexpected. *iscR* C104S has been previously reported [126], [182]. The mutation is in IscR's own Fe-S binding site, which causes it to maintain an unbound state that should de-repress Isc and activate Suf

[181], [182]. We observe a strong upregulation of Isc in these strains, with more modest increases in Suf iModulon activity. The other most common mutation, *iscR* V55L, seems to downregulate Isc while also keeping Suf near basal levels. Given that ROS stress induces Fe-S cluster damage and Suf is significantly better at handling ROS stress [180], we would initially expect mutations which upregulate Suf to be more effective under the ALE conditions and therefore be enriched in these strains. We only see one mutation, *iscR* V87A, which seems to achieve that.

One possible explanation for this unexpected outcome is that the proteomic cost of the systems, particularly Suf, selects against strains which allocate too many resources towards Fe-S synthesis; this explanation has been modeled in a ME flux balance analysis [126]. However, another possibility relates to the control of electron flux described in **Sections 4.2.9–4.2.11**: many redox enzymes, including some in respiration and the TCA cycle, contain Fe-S clusters [183]. Damage to these enzymes by high ROS slows oxidative metabolism. This would charge fewer electron carriers and therefore slow the PQ cycle, allowing the cell to recover. It would therefore be better to express less Suf so that Fe-S synthesis would remain sensitive to ROS — using Isc or less of both systems would strengthen the coupling between ROS and respiration as a means of controlling the PQ cycle (**Figure B.1E**). Thus, like the *fur* P18T mutation (**Section 4.2.6.2**), this mutation enables a negative feedback loop, which aids in slowing oxidative metabolism and PQ cycling when stress is high (**Figure B.1E**).

While discussing Fe-S clusters, it is also worth noting that every strain mutated the putative Fe-S cluster repair gene *ygfZ*. This provides evidence for its putative role in Fe-S cluster homeostasis and motivates further study.

4.2.6.4 SoxS Activity Does Not Change With Adaptation, Suggesting a Key Role in Stress Readiness

Interestingly, there was a lack of mutations affecting *soxS*, the regulator of processes that remove the ROS superoxide[184]. SoxS iModulon activity is highly correlated with PQ in the starting and all evolved strains (**Figure B.1F**; Pearson R = 0.72, p = $5.5*10^{-15}$). The lack of mutations suggests that ROS readiness is preserved by using wild-type *soxS*. Despite the downregulation of general stress responses (**Section 4.2.8**), this specific stress response is preserved.

4.2.7 An Unexpected Mutation in *pitA* Regulates Motility

A frameshift in the phosphate transporter *pitA* led to a motile phenotype. This mutation occurred in 1_0 and its derivatives, and these strains also exhibited strong activation of motility-associated iModulons such as FliA (**Figure 4.5H**). There is no obvious connection between phosphate transport and motility, and the mutated strains were likely able to use the other phosphate transport system, *pstABCS*, to meet their phosphorus needs [185]. Interestingly, the 3_0 strain deleted *pitA* as part of Del-1 (**Figure 4.4B**), and it did not exhibit the motility phenotype. Thus, to understand this mutation, we generated two new strains: 0_0::*pitA** and 1_0::*pitA*, which added the mutation on its own to the starting strain and removed it in favor of the original *pitA* sequence in the evolved strain, respectively. We found that the mutation provided a growth advantage under PQ stress (**Figure 4.5I**). We also transcriptomically profiled the strains under the same conditions used for our other strains, and found that, particularly under PQ stress, the mutation exclusively perturbs the motility iModulons (**Figure 4.5J**). The change to the transcriptome was also reflected in the phenotype, as the mutant strains swarmed on agar plates while the wild-type *pitA* strains did not (**Figure 4.5K, B.2**). The detailed mechanism of action linking the *pitA* mutant to motility remains to be elucidated.

An upregulation of anaerobic iModulons such as Fnr-3 in the ALE *pitA* mutants (**Figure B.1G-I**) suggests a possible benefit for motility, in that it may be correlated with beneficial fermentation phenotypes discussed later (**Section 4.2.9–4.2.11**). The gene *aer*, which is upregulated as part of the FliA iModulon, mediates aerotaxis and would therefore allow cells to swim away from locally high concentrations of ROS [186], [187]. In addition to its role in chemotaxis, *aer* helps to upregulate the Entner-Doudoroff pathway and anaerobic metabolism [188], a tendency which can be observed in the iModulon activities of our strains. Each of the anaerobic iModulons, Fnr-3 in particular, is slightly upregulated by the strains with *pitA** (**Figure B.1G-I**). An increase in anaerobic metabolism would help to prevent PQ cycling as described in **Section 4.2.9** and **4.2.10**. Thus, a decrease in oxidative metabolism is also achieved by the cells through this very non-conventional mechanism. An added benefit may lie in the expression of *fliZ*, a member of the FliA iModulon, which is known to antagonize RpoS and would therefore be expected to promote growth [189].

This section illustrates the usefulness of our multilevel approach. After connecting mutations to their effects and predicting causes for DiMAs, we were left with an orphan mutation (*pitA*) and an unexplained DiMA (FliA). We predicted that the mutation caused the DiMA, and then we generated new strains to validate the prediction. The recapitulation of the expected iModulon change and swarming phenotype lends credibility to the iModulon method of elucidating mutational effects.

4.2.8 Shifting from Stress to Growth Explains Activity of Several iModulons

Regulatory iModulons can be used not only to understand the direct effects of mutations as described above, but also effects of changes to the processes that TFs sense. We have divided these types of changes in the PQ tolerant strains into two categories: those that respond to stress and growth (21% of the variance in the transcriptome; **Figure 4.2H**), and metabolic changes (10%). In this section we describe the former.

4.2.8.1 The ‘Fear-Greed Tradeoff’ Shifts Towards Greed with Evolution

An important global tradeoff in the *E. coli* transcriptome is between growth and general stress readiness, which is governed by complex regulation [190], [191]. We previously identified a ‘fear-greed tradeoff’ between the RpoS and Translation iModulons, in which the activity levels of the two iModulons have a negative correlation; faster growing cells exhibit low RpoS and high Translation activity [19], [40], [43], [192], [193]. The starting strain without stress is ‘greedy’, but it becomes ‘fearful’ upon addition of PQ, as expected (**Figure 4.6A-B**). The evolved strains, on the other hand, largely remain ‘greedy’ in the presence of PQ; they strongly downregulate RpoS (**Figure 4.6A**) and have higher translation activity than the stressed starting strain (**Figure 4.6B**). Translation activity is decreased relative to the starting strain in the absence of PQ, likely because of tradeoffs towards ROS stress readiness in the tolerized phenotype.

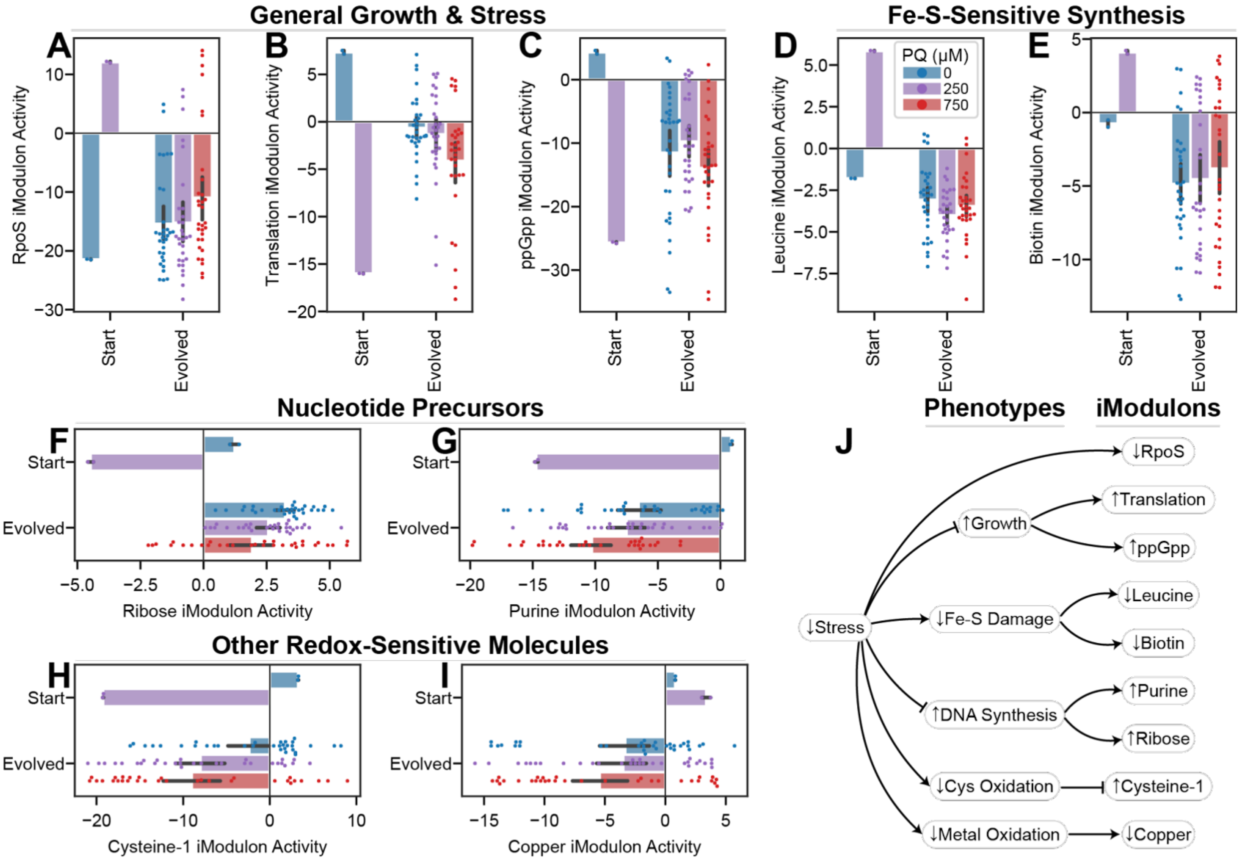


Figure 4.6: Changes to stress and growth explain the changes to activity in several iModulons. Mean iModulon activities ± 95% confidence interval; all plots use the legend in (D). P-values are false discovery rate corrected p-values from a comparison of stressed transcriptomes (250 and/or 750 µM PQ) between 0_0 and evolved strains. **(A)** RpoS activity, the general stress response, is downregulated ($p = 0.017$). **(B)** The Translation iModulon, ribosomes and translation machinery, is upregulated ($p = 0.023$). **(C)** The ppGpp iModulon, a large iModulon with many growth-related functions, follows a similar pattern to the Translation iModulon ($p = 0.027$). **(D)** The Leucine iModulon, which responds to leucine concentrations downstream of an Fe-S-dependent synthesis pathway, is downregulated after evolution, suggesting improved Fe-S metabolism ($p = 0.0017$). **(E)** The Biotin iModulon is downregulated after evolution. Biotin also depends on Fe-S-dependent synthesis ($q = 0.017$). **(F-I)** Ribose ($p = 0.011$), Purine ($p = 0.036$), Cysteine-1 ($p = 0.025$), and Copper ($p = 0.034$) iModulon activities behave differently in starting and evolved strains. **(J)** Knowledge graph connecting decreased oxidative stress to each of the iModulon changes shown.

The ppGpp iModulon contains a large set of growth-related genes regulated by the master regulator ppGpp [194]. It follows a similar pattern to the Translation iModulon, suggesting that ppGpp concentrations decline after evolution (**Figure 4.6C**). In addition to the Translation and ppGpp iModulons, a few other differentially activated iModulons with more specific functions are also likely to be responding to ppGpp levels, including the Nucleotide Stress, Glutarate, Efflux Pump, and Biofilm iModulons.

Despite the presence of stressors and the activation of specific ROS responses OxyR and SoxS (**Sections 4.2.6.1** and **4.2.6.4**), the general stress response is not activated in the evolved strains. There are two likely reasons for this: the stress signals are downregulated by the success of the evolved strategies of PQ tolerance, and the growth-inhibiting effects of RpoS have selected against strains with high RpoS activity. This work agrees with previous findings that ALE shifts allocation toward 'greed' [19], [40], [192]. The decoupling of the ROS and general stress responses makes these strains ROS-response specialists, constituting a valuable adaptation strategy.

4.2.8.2 Functional Iron-Sulfur Clusters Downregulate Leucine and Biotin Synthesis

Two DiMAs reflect a decrease in oxidative damage by sensing Fe-S-dependent metabolites. The Leucine iModulon (**Figure 4.6D**) encodes the leucine biosynthesis pathway, which requires an Fe-S cluster and other metal-dependent enzymes that are sensitive to oxidative stress [195]. Leucine feeds back to inhibit the iModulon's expression [196]. In the starting strain with PQ, oxidative damage likely leads to a decrease in leucine concentrations and an upregulation of the iModulon. By contrast, the evolved strains experience less stress, protect their Fe-S clusters, and therefore exhibit low Leucine iModulon activity. Similarly, the Biotin iModulon (**Figure 4.6E**) uses an Fe-S cluster in BioB to synthesize biotin [197], which then controls iModulon activity via regulation by BirA [198].

4.2.8.3 Increased Demand for Nucleotides underlies Purine and Ribose Synthesis

Ribose concentrations are sensed by RbsR [199], which represses the Ribose iModulon in its presence. Ribose is produced as part of the pentose phosphate pathway (PPP), which is the primary pathway for producing NADPH to detoxify ROS. Upon initiation of oxidative stress, PPP flux increases, producing ribose [200]. Oxidative stress also slows growth and DNA synthesis, which will decrease ribose utilization. We therefore expect an increase in ribose concentrations in the starting strain upon PQ stress, which is observed as a decrease in Ribose iModulon activity (**Figure 4.6F**). In the evolved strains, flux shifts towards glycolysis and away from the PPP, producing less ribose. They also synthesize more DNA to support faster growth, using ribose. Therefore, Ribose iModulon activity increases relative to the starting strain, while still exhibiting a negative correlation with PQ.

The Purine iModulon is regulated by PurR and ppGpp, and its activation pattern in our samples (**Figure 4.6G**) mirrors that of the Translation and ppGpp iModulons (**Figure 4.6B–C**). This activation may be explained by direct action by ppGpp, or via PurR, which represses these genes in the presence of hypoxanthine or guanine [201]. The faster growing evolved strains would perform more DNA replication and RNA synthesis, and therefore require purine synthesis, depleting the metabolites which are sensed by PurR and de-repressing the iModulon.

4.2.8.4 Cysteine and Copper iModulons Sense the Oxidation State of the Cell

Changes in Cysteine-1 iModulon activities may be explained by increased ROS readiness and subsequent improvement in amino acid homeostasis (**Figure 4.6I**). This iModulon is regulated by CysB, which can be inhibited by cystine and other oxidized sulfur compounds [202], [203]. Cysteine is very easily oxidized [204], [205], which may explain the dramatic downregulation of the iModulon upon PQ addition in the starting strain. The evolved strains with PQ have significantly higher Cysteine-1 activity compared to the parent strain with PQ, due to the success of their tolerization strategies.

The Copper iModulon, which contains copper efflux genes regulated by CueR, CusR, and HprR, is downregulated in the evolved strains (**Figure 4.6K**). Copper is redox-sensitive, and its efflux depends on the proton-motive force (PMF) or ATP [206]. It is also an important cofactor for various enzymes, including the superoxide dismutase *sodC* [207]. Oxidative damage should decrease the PMF and ATP concentrations and alter the copper redox state, which would explain the iModulon's upregulation in the stressed starting strain. The evolved strains downregulate this iModulon, reflecting improvements in metal homeostasis resulting from ROS tolerization.

Thus, iModulons measure the entire sensory output of the TRN and allow us to mine the transcriptome for insights into many cellular processes. Because we also have an understanding of the stress phenotype of the cells, we predicted reasons for a large fraction of transcriptional alterations. This approach would be useful to any researcher seeking to enumerate phenotypic alterations in novel strains using only RNAseq data as a guide.

4.2.9 Mutating Central and Energy Metabolism Genes Decreases PQ Cycling

We now turn to metabolism, which adds a fourth layer to our analysis and involves a complex interplay of effects from each level (**Figure 4.3A**). We show that enzyme mutations can suggest tolerance strategies, and then ME modeling can validate them. Finally, iModulon analysis can reveal how those strategies are organized and regulated by the cell.

4.2.9.1 Loss-of-function TCA Cycle Mutations Were Commonly Acquired

The main metabolic mutations occur in the tricarboxylic acid (TCA) cycle. The second gene to mutate in all strains was *aceE* (**Figure 4.2B,D**). *aceE* encodes a subunit of pyruvate dehydrogenase (PDH), the entry point into the TCA cycle. *gltA, sucA,* and *icd* also mutate often, with *icd* being affected by e14 deletion and SNPs [208] (**Figure 4.7A**). These mutations would likely decrease the function of the enzymes, thus decreasing TCA cycle flux and production of NADH. These mutations suggest a tolerance benefit to decreasing NADH production. The likely reason for this benefit is that PQ uses electrons from NAD(P)H to reduce oxygen and generate stress [209]–[211]. These mutations would decrease the available electrons to the PQ cycle and prevent stress generation. To decrease oxidative stress from PQ, the evolved strains perform less oxidative metabolism. The Fe-S and motility mechanisms (**Sections 4.2.6.3** and **4.2.7**) also shift strains away from NADH production.

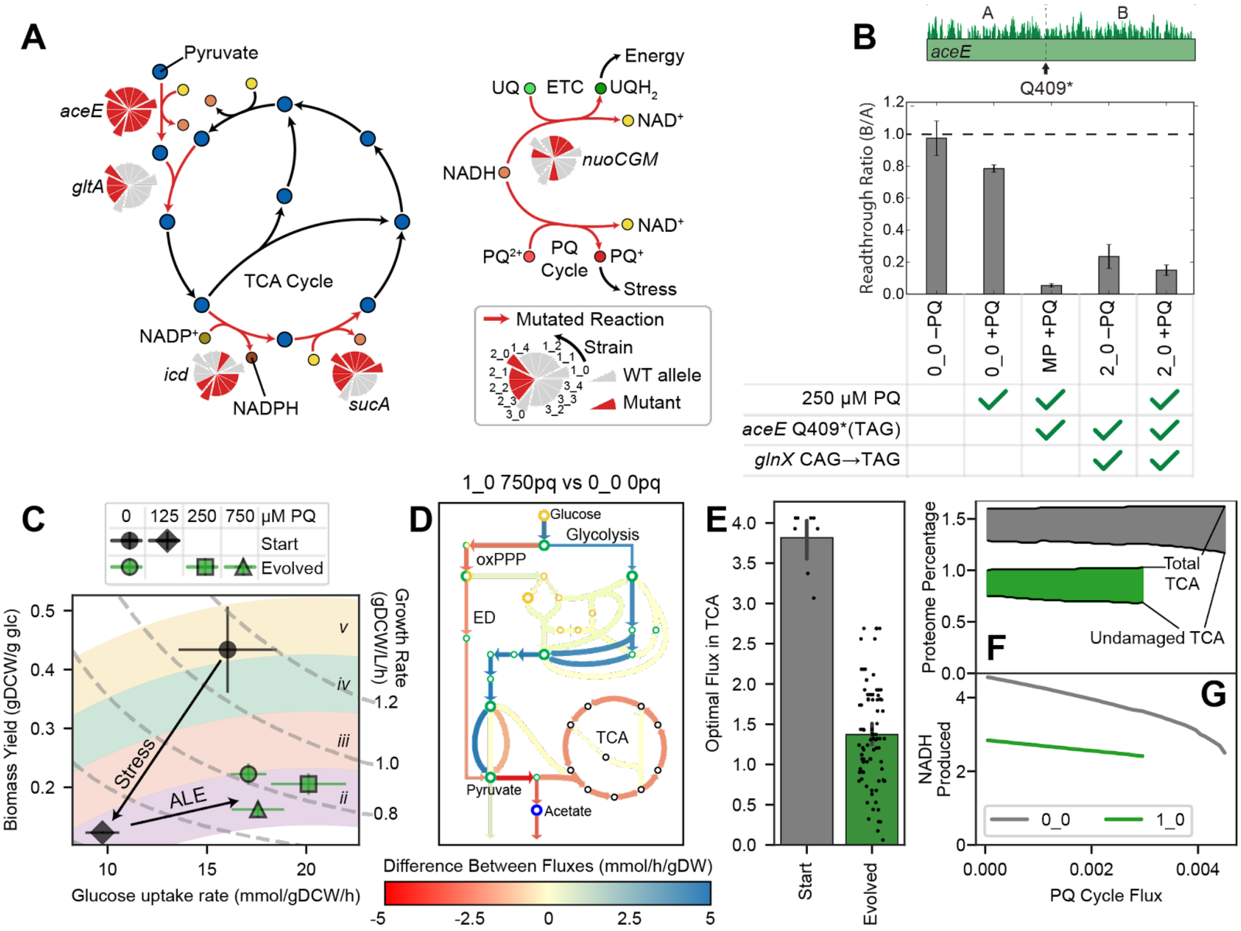


Figure 4.7: Mutations drive metabolic rerouting toward fermentation to avoid PQ cycling by decreasing NADH availability. **(A)** Simplified metabolic map of the TCA cycle and fate of NADH. Reactions catalyzed by mutated enzymes are shown in red and labeled with a pie chart indicating which strains have a wild-type (WT) or mutant allele. First generation strains in the pie chart protrude, with their descendants following them counter-clockwise. **(B)** Ribosome readthrough ratio in *aceE* from ribosome profiling, means ± standard deviation. The ratio B/A is the fraction of ribosomes bound downstream (B) vs. upstream (A) of the early amber stop codon (TAG) in aceE. The midpoint (MP) strain has *aceE* Q409* with WT *glnX*, whereas the 2_0 strain has both *aceE* Q409* and the *glnX* anticodon mutation that enables ribosomes to read through the amber stop codon. In evolved strains such as 2_0, PDH levels are decreased but not zero. **(C)** Aero-type plot [212] computed from measured growth rates and glucose uptake rates, where points represent means ± SEM, with constant growth rate isoclines. Colored regions labeled with roman numerals are aero-type regions as defined previously [150]. Cells switch to a lower aero-type with PQ and increase their glucose uptake after evolution. **(D)** Flux differences from the OxidizeME model, comparing the starting strain with no PQ and a representative evolved strain at high PQ. Model was constrained by growth rate, glucose uptake rate, and RNAseq data (**Figure B.4**). **(E)** Each point represents a TCA cycle reaction in the constrained OxidizeME models; models of evolved strains predict lower TCA cycle fluxes. **(F-G)** OxidizeME model results in mmol/gDCW/h for 0_0 and 1_0, constrained by growth rate, glucose uptake rate, and RNA expression. **(F)** As PQ cycle flux increases, the damaged fraction (filled in) of the TCA cycle increases. **(G)** NADH production decreases with PQ, but is more sensitive in 0_0. 0_0 can also carry more PQ cycle flux.

Loss of function (LOF) mutations in the TCA cycle come with a cost, since those pathways are the primary energy source for aerobic cells. Indeed, the evolved strains have decreased growth and translational activity under no stress relative to the starting strain, probably for this reason (**Figure 4.2C, 4.6B**). During ALE, the strains must therefore balance a tradeoff: generate enough NADH to grow and repair themselves, but not so much as to over-empower the PQ cycle.

We summarize all metabolic mutations in **Figure B.3**.

4.2.9.2 Synergistic Mutations in *aceE* and *glnX* Balance a Tradeoff Between Energy and Stress

The growth/stress tradeoff of the TCA cycle is embodied by interactions between two mutations, which both occurred in both the 1_0 and 2_0 strains. First, *aceE* acquired a C→T nonsense SNP, creating an amber stop codon [213]: Q791* in 1_0 and Q409* in 2_0. This mutation inactivated PDH and likely significantly decreased flux into the TCA cycle. While effective early in the evolution at decreasing PQ cycling, the change was extremely damaging. Interestingly, both 1_0 and 2_0 later acquired the same C→T SNP in the anticodon of the glutamine tRNA *glnX* [214]. This second change enabled the mutant *glnX* to read through the initial *aceE* truncation, allowing for some functional PDH to be translated and utilized for energy generation. Due to competition between stop codon release factors and *glnX*, functional *aceE* translation would not return to wild type levels [215], but rather find an intermediate level which balanced the tradeoff.

We quantified the above relationship using ribosome profiling (**Figure 4.7C**). By measuring the fraction of ribosomes bound to the sequence before and after the truncating SNP, we demonstrated the near complete deactivation of *aceE* translation in the midpoint strain. In the 2_0 strain with both the *aceE* and *glnX* mutations, translation was partially restored (to a ratio of 0.23±0.08). Thus, synergy between these two mutations brokered a compromise between the energy and stress-generating effects of TCA cycle flux.

The 3_0 strain acquired a frameshift 1 bp deletion in *aceE* instead of the nonsense SNP. This meant that it could not employ a similar strategy to 1_0 and 2_0. However, two of its second generation derivative strains (3_1 and 3_3) had insertions at or near the deletion (**Figure B.3**), which may have served a similar purpose in re-increasing PDH levels.

4.2.9.3 Mutations in NADH Utilization Genes Suggest PQ Diaphorase Activity

In addition to the NADH production-related mutations described in **Sections 4.2.9.1–4.2.9.2**, we also observe NADH utilization-related mutations. Five strains acquired unique mutations in *nuoC, nuoG,* and *nuoM* of the NADH dehydrogenase complex (NDH-1). A 40 bp deletion within *nuoG* appears to induce early termination of transcription, since genes downstream of it are captured by the NDH-1 iModulon and strongly downregulated in the strain with the deletion (**Figure B.1J**). Note that another strain deleted 123 bp in a nearby region of the same gene, but we do not observe early termination in that strain. The prevalence of these mutations suggests a benefit to NDH-1 LOF under PQ conditions.

Cellular enzymes which catalyze PQ reduction are called PQ diaphorases, and three have been identified in *E. coli* by past studies [210], [216]. Those studies suggested that NADPH plays a larger role than NADH, but our mutations preferentially affect NADH production and NDH-1. It is possible that transhydrogenases first convert NADH to NADPH [217] prior to the PQ cycle. Alternatively, NDH-1 and other mutated NADH reductases from this study (e.g. *cyoB, ubiF, torZ,* and *trxC;* Figure B.3) ought to be considered as potential PQ diaphorases. Though NDH-1 has not been implicated in PQ cycling in *E. coli*, this phenomenon has been observed in mammals [209], [211].

4.2.10 Metabolic Rewiring Towards a Lower Aero-Type Decreases PQ Sensitivity and Flux in Evolved Strains

4.2.10.1 The Evolved Strains Decrease their Aero-Type

We quantified glucose uptake for each strain at various PQ levels, and generated a plot comparing biomass yield per gram of glucose to the glucose uptake rate (**Figure 4.7C**). This rate-yield plane has been characterized in past studies [150], [212], which revealed distinct energy generation strategies (aero-types) for each position in the plane. Samples with high biomass yields are in the highest aero-type (aero-type *v*), which represents efficient aerobic growth, whereas lower aero-types are associated with lower aerobicity and secretion of organic acids. The higher aero-types pump more protons across the inner membrane than the lower aerotypes [150].

In **Figure 4.7C**, we observe a switch to a lower aero-type in the starting strain upon PQ exposure, since ROS damage decreases growth rate and particularly damages respiration. In the evolved strains, the lower aero-type is maintained even when no PQ is present. The aero-type change is likely due to the TCA cycle-related mutations, which we predicted would decrease respiration. However, the evolved samples also shift rightward, increasing their glucose uptake and total metabolic flux, enabling them to maintain growth under stress. Their position in the plane doesn't vary much with PQ concentration, indicating decreased sensitivity.

4.2.10.2 Metabolism and Expression Modeling Demonstrates Decreased PQ Cycle Flux and Sensitivity

To characterize metabolism *in silico*, we used OxidizeME, a genome-scale computational model of *E. coli* metabolism and expression (ME) which incorporates ROS stress effects [126]. We constrained the model using each strain's growth rate, glucose uptake rate, and RNA expression, then simulated optimal steady states (**Figure 4.7D, B.4**). Though we did not attempt to simulate the effects of mutations on the reaction rates, the optimal flux distributions in the evolved strains showed decreases in TCA cycle flux (**Figure 4.7E**), consistent with the predicted effects of the mutations.

In the absence of experimental methods for directly measuring PQ cycle flux, we computationally assessed the consequences of PQ cycle flux by varying it for the starting strain and a representative evolved strain (**Figure 4.7F-G**). Though total proteomic allocation to the TCA cycle was constrained to match the RNA expression, ROS damage to the Fe-S clusters in *acnA*, *fumAB*, and *sdhABCD* led to decreasing functional proteome fractions (**Figure 4.7F**). The starting strain relied more heavily on the TCA cycle; this made it more sensitive to PQ, as evidenced by the steeper slope in NADH production (**Figure 4.7G**). The starting strain was also able to grow at higher PQ fluxes, which is inefficient and exacerbates stress. Thus, tolerization both decreases sensitivity to lower PQ fluxes and prevents a steady state with high PQ flux.

The genome-scale OxidizeME model integrates the individual cellular processes and RNA expression changes which adjust the phenotype, and it elucidates key systems level tolerization strategies. Its results match expectations from mutational analysis.

4.2.11 iModulon Activities Shift Tolerant Strains Towards Anaerobic Metabolism and Glycolysis

4.2.11.1 Increased Fermentation Is Induced by ArcA and Fnr

Finally, we discuss iModulons which regulate the metabolic rerouting presented above. The cellular oxidation state is sensed and regulated by ArcA and Fnr [218], whose iModulons are differentially activated in the evolved strains (**Figure 4.8A-D**). Both TFs sense redox balance, which shifts towards reduction in the evolved strains due to the successful tolerization: ArcA represses when the electron transport chain is in a reduced state [219], whereas Fnr repression ceases when Fe-S clusters are intact [220] (**Figure 4.8E**). These transcriptional changes shift from aerobic respiration genes toward anaerobic fermentation genes [218] (despite the aerobic ALE conditions). This strategy maintains a lower aero-type and decreases reliance on NADH. Thus, this mechanism reinforces the decreased reliance on the TCA cycle brought on by the mutations, ultimately slowing PQ cycling.

ArcA is part of the ArcAB two-component system, which senses the ratio of reduced to oxidized quinones in the ETC [219]. In the starting strain, oxidative stress from PQ shifts this ratio toward oxidation, causing ArcAB to be less active and derepress the ArcA iModulon. As strains evolve, they experience less oxidative stress due to their transport and TCA cycle mutations. This lowered stress leads to a more reduced quinone pool, an increase in ArcAB activity, and repression of the ArcA iModulon (**Figure 4.8A**). The ArcA iModulon contains aerobic growth genes such as oxidoreductases and cytochromes, so its repression will encourage anaerobic metabolism, fermentation, and a decreased reliance on NADH.

Fnr senses oxygen levels via oxidative damage to its Fe-S cluster and activates anaerobic metabolism genes when the cluster is intact [220], [221]. Its regulon is captured by three iModulons, whose activities behave similarly in this study (**Figure 4.8B-D**). The decrease in oxidative stress, as well as the success of iron-related mutations, help to maintain more active Fnr and therefore upregulate this iModulon.

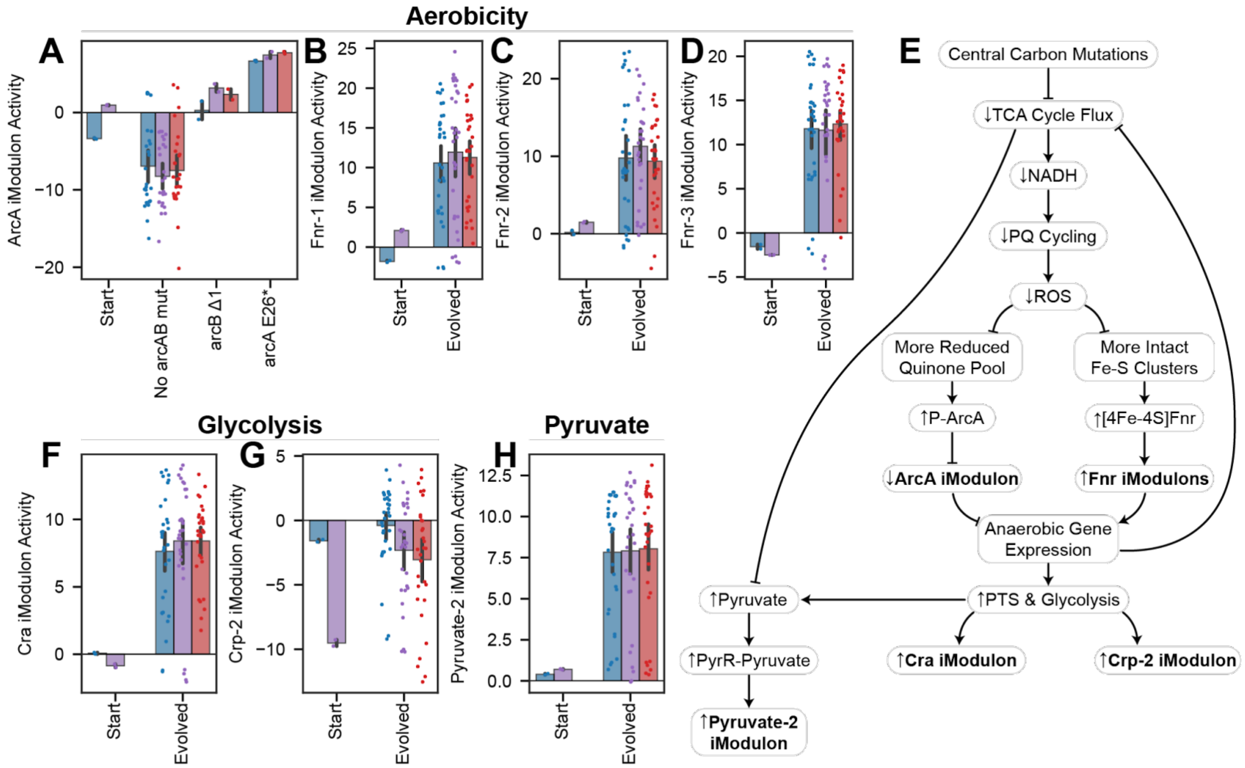


Figure 4.8: Mutations and iModulon reallocation drive metabolic rerouting toward fermentation to avoid PQ cycling. Bars indicate mean iModulon activities ±95% confidence interval. **(A)** ArcA iModulon activities are mostly decreased after evolution, except in the case of mutations to *arcAB* (p = 0.035). ArcA contains aerobic metabolism genes. **(B-D)** Fnr controls three iModulons with anaerobic metabolism genes, all of which are upregulated (p = 0.034, 0.030, 0.023). **(E)** Knowledge graph describing changes in the evolved strains connecting central carbon mutations to anaerobic and glycolytic gene expression, which decreases TCA cycle flux and ROS generation. **(F)** The Cra iModulon, which contains glycolytic genes that are repressed by Cra, is upregulated (p = 0.017). **(G)** The Crp-2 iModulon, which controls phosphotransferase systems, is upregulated (p = 0.022). (H) The Pyruvate-2 iModulon is upregulated (p = 0.012).

There are two strains which have mutations in the ArcAB two-component system, affecting ArcA iModulon activity. A frameshift in the sensor kinase *arcB* in 3_1 has a moderate derepressing effect, and an early stop in *arcA* in 1_1 had a stronger derepressing effect (**Figure 4.8A**). These two strains are an exception which appear to have struck a different balance in the growth/stress generation tradeoff compared to the other evolved strains. They express aerobic metabolism genes as well as the Fnr-activated anaerobic fermentation genes, which would enable them to use more energy producing pathways but could also exacerbate stress generation.

4.2.11.2 Glycolytic Flux Increases Due to the Action of Cra and Crp

To meet energy needs with lower respiration, the cells increased their glycolytic activity, a change which is described by two DiMAs. Cra iModulon activity increases, indicating an increase in glycolytic flux (**Figure 4.8F**). Similarly, the Crp-2 iModulon returns to unstressed or intermediate levels in the evolved strains, which indicates a more active phosphotransfer system (**Figure 4.8G**). This transcriptomic change matches the rightward shift in the aero-type plot (**Figure 4.7C**).

The Cra iModulon captures a set of genes of glycolysis and carbohydrate catabolism genes which are repressed by Cra [222], [223]. Cra regulates these genes by acting as a flux sensor for glycolysis, since their suppression is activated by fructose-1,6-bisphosphate [224]. We observe an increase in Cra iModulon activity in the evolved strains (**Figure 4.8F**), which both indicates and positively regulates an increase in glycolytic flux.

The Crp-2 iModulon contains mostly phosphotransfer (PTS) system genes which are activated by the master regulator Crp [225]. Crp responds to cAMP levels in a biphasic manner, and cAMP levels themselves have complex regulation [226]. We observe a strong downregulation of the Crp-2 iModulon in the stressed starting strain, but a return to unstressed or intermediate levels in the evolved strains (**Figure 4.8G**). This change is consistent with a return to homeostasis, and may indicate a more active PTS, higher glucose uptake, and increase in ATP concentrations after evolution.

4.2.11.3 Pyruvate Accumulation is Sensed by Exometabolomics and the Pyruvate-2 iModulon

The LOF mutations in PDH and the TCA cycle should increase intracellular pyruvate concentrations, since pyruvate is the initial substrate for those reactions. The Pyruvate-2 iModulon is regulated by PyrR, which can sense pyruvate concentrations [227]. Pyruvate-2 activity increases in the evolved strains (**Figure 4.8H**), which is consistent with this prediction. We also observe pyruvate secretion at high PQ levels (**Figure B.1K**), probably due to the oxidative damage to PDH and the TCA cycle causing so much pyruvate accumulation that it must be secreted.

In **Sections 4.2.9–4.2.11**, we showed that mutations and iModulon activity adjustments work together to enforce a low aero-type, PQ-tolerant metabolic network. The PQ tolerance stems from a decreased reliance on the TCA cycle and decreased NADH production, which leads to a metabolic network that supports less total PQ cycling and makes the system less sensitive to small amounts of PQ cycling. It is often difficult to interpret biological systems when genes, gene expression, and metabolic flux are all changing, but our multilevel interoperable approach using mutational analysis, iModulon activity changes, and genome-scale modeling produced a consistent and comprehensive interpretation of multiple data types.

4.3 Discussion

In this chapter, we combined ALE with a detailed, systems-level transcriptomic analysis to comprehensively reveal mechanisms underlying PQ tolerance. The approach spanned four levels of analysis (**Figure 4.3A**): (i) genetic alterations and their predicted effects, (ii) transcriptomic adaptations along with up- and downstream inferences about their regulatory causes and physiological impact, (iii) metabolic fluxes calculated from genome-scale metabolic models, and (iv) phenotypic changes such as swarming motility. We found iModulon analysis of the transcriptome to be particularly revealing, as the TF activities could be readily quantified and utilized to infer a wealth of information about the phenotypic state. By combining these approaches into a coherent set of tolerization strategies, we presented a summary of the systems biology of paraquat tolerance.

The evolved strains characterized herein achieved high tolerance through several mechanisms (**Figure 4.3B**). They promoted efflux of PQ via *emrE* segmental amplification, and precluded influx by mutating or deleting various other transporters. Inside the cells, PQ failed to generate as much ROS due to LOF mutations in and downregulation of NADH-producing pathways. To compensate for the decreased biomass yield of their metabolism, the cells increased glucose uptake and glycolytic flux. Since ROS interact with iron, some strains modified iron regulation via TF mutations that curtailed these systems when stress was high. These mutational and metabolic strategies led to a decrease in stress, which was sensed by the TRN and shifted various regulators toward faster growth.

The impact of this study is threefold. (i) We present biological insights of wide interest to researchers, including the growth/stress tradeoff of redox metabolism, the use of Fe-S clusters as a brake on iron uptake and metabolism, and novel interactions such as those between *pitA* and motility and between *aceE* and *glnX*. (ii) Acquired mutations and iModulon activities can become design variables for strain engineering, which frequently seeks to mitigate oxidative stress for bioproduction applications. (iii) We demonstrate an approach that utilizes iModulons to reveal a novel integrated perspective on adaptation to stress by understanding transcriptomic allocation. This approach will be reused in **Chapter 5** to understand adaptation to a different type of stress.

Future studies should integrate additional data types into this framework. For instance, proteomics, endo-metabolomics, and chromatin immunoprecipitation of key TFs would be able to test various aspects of these hypotheses, better constrain models, and potentially uncover new insights. In addition, we encourage focused studies which characterize the mechanisms proposed here in greater detail.

Taken together, our results elucidate the systems biology of paraquat tolerization using genome-scale datasets, computational models, and detailed literature review. Given the falling cost of RNAseq, development of laboratory evolution, and the availability of the pipeline developed here, we can expect that the systems biology of an increasing number of cellular functions and adaptations will be revealed.

4.4 Methods

4.4.1 Data and Code Availability

RNAseq data have been deposited to GEO and are publicly available as of the date of publication, under accession numbers GSE134256 and GSE221314. DNAseq data are available from aledb.org under the project "ROS". iModulons and related data are available from iModulonDB.org under the dataset "*E. coli* PRECISE-1K".

All original code and data to generate figures are available at github.com/SBRG/ROS-ALE, which also links to the alignment, ICA, and iModulon analysis workflows [128]. It has been deposited at Zenodo and is publicly available [228]. The DOI is 10.5281/zenodo.7449004.

4.4.2 Microbial Strains and Culture Conditions

The starting strain (0_0) was an MG1655 K-12 *E. coli* strain which had been evolved for optimal growth on glucose as a carbon source in M9 minimal media[162]. Mutations for the evolved strains are listed on aledb.org.

Strains were grown overnight in M9 minimal media with 0.4% w/v glucose as a carbon source. Fresh media was inoculated with the overnight culture at an initial 600 nm optical density (OD) of 0.025. Cultures were aerated with a stir bar at 1100 rpm in a water bath maintained at 37°C until OD reached 0.5. 50 mM PQ was added to reach the desired concentration in stressed flasks. After 20 minutes, samples were harvested for transcriptomics or ribosome profiling.

4.4.3 Adaptive Laboratory Evolution

ALE was performed using a similar protocol to Mohamad *et al.* 2017 [229]. Parallel cultures were started in M9 minimal medium by inoculation from isolated colonies. Evolution was performed in an automated platform with 15 mL working volume aerobic cultures maintained at 37°C and magnetically stirred at 1100 rpm. Growth was monitored by periodic measurement of the 600 nm OD on a Tecan Sunrise microplate reader, and cultures were passaged to fresh medium during exponential cell growth at an OD of approximately 0.3. Growth rates were determined for each batch by linear regression of ln(OD) versus time. At the time of passage, PQ concentration in the fresh medium batch was automatically increased if a growth rate of 0.08 h^{-1} had been met for 3 consecutive flasks. Samples were saved throughout the experiment by mixing equal parts culture and 50% v/v glycerol and storing at -80°C.

4.4.4 DNA Sequencing and Mutation Calling

DNA was isolated as described [230]. Total DNA was sampled from an overnight culture and immediately centrifuged for 5 min at 8,000 rpm. The supernatant was decanted, and the cell pellet was frozen at -80°C. Genomic DNA was isolated using a Quick-DNA Fungal/Bacterial Microprep Kit (Zymo Research) following the manufacturer’s protocol, including treatment with RNase A. Resequencing libraries were prepared using a Kapa Hyper Plus Kit (Roche Diagnostics) following the manufacturer’s protocol. Libraries were run on HiSeq and/or NextSeq (Illumina).

Sequencing reads were filtered and trimmed using AfterQC version 0.9.7 [231]. We mapped reads to the *E. coli* K-12 MG1655 reference genome (NC_00913.3) using the breseq pipeline version 0.33.1 [232]. Mutation analysis was performed using ALEdb [146].

4.4.5 Physiological Characterization

Growth curves and exometabolomic samples were generated by inoculating cells from an overnight culture to a low OD using the same conditions as the ALE. For each strain, we started with 0 PQ. OD measurements and samples were taken at various time points until stationary phase was reached. We then passaged the cells into a new flask, stepped up the PQ concentration, and characterized the next curve, for concentrations 125, 250, 500, 750, 1500, and 2500 µM. We stopped if growth was not observed after 48 hours. For each flask, growth rates were determined by linear regression of ln(OD) versus time in the early exponential part of the curve.

We took cell culture samples at the same time as OD measurements for the starting strain at 0 and 125 µM PQ, and for the evolved strains at 0, 250, and 750 µM PQ. Samples were sterile filtered, and extracellular by-products were determined by high pressure liquid chromatography (HPLC). The filtrate was injected into an HPLC column (Aminex HPX-87H 125-0140). The concentrations of the detected compounds were determined by comparison to a normalized curve of known concentrations. Substrate uptake and secretion rates in the early exponential growth phase were calculated from the product of the growth rate and the slope from a linear regression of the grams dry weight (gDW) versus the substrate concentration. The biomass yield was calculated as the quotient of the growth rate and the glucose uptake rates during the exponential growth phase.

4.4.6 RNA Sequencing

3 mL of induced culture was added to 6 mL of RNAProtect Bacteria Reagent (Qiagen) and vortexed, then left at room temperature to incubate for 5 minutes. Cells were pelleted, resuspended in 400 µL elution buffer, and then split into two tubes with one kept as a spare. One pellet was then lysed enzymatically with addition of lysozyme, proteinase-K, and 20% SDS. SUPERase-In was added to maintain the integrity of the RNA. RNA isolation was then performed according to the RNeasy Mini Kit (Qiagen) protocol. rRNA was depleted using the Ribo-Zero rRNA Removal Kit for gram negative

bacteria according to the protocol. Libraries were constructed for paired-end sequencing using a KAPA RNAseq Library Preparation kit. Reads were sequenced on the Illumina NextSeq platform.

As part of the PRECISE-1K dataset [140], transcriptomic reads were mapped using our pipeline (https://github.com/avsastry/modulome-workflow) [128] and run on Amazon Web Services Batch. First, raw read trimming was performed using Trim Galore with default options, followed by FastQC on the trimmed reads. Next, reads were aligned to the *E. coli* K-12 MG1655 reference genome (NC_000913.3) using Bowtie [233]. The read direction was inferred using RSeQC [234]. Read counts were generated using featureCounts [235]. All quality control metrics were compiled using MultiQC [236]. Finally, the expression dataset was reported in units of log-transformed transcripts per million (log(TPM+1)).

All included samples passed rigorous quality control, with "high-quality" defined as (i) passing the following FastQC checks: *per_base_sequence_quality, per_sequence_quality_scores, per_base_n_content, adaptor content;* (ii) having at least 500,000 reads mapped to the coding sequences of the reference genome (NC_000913.3); (iii) not being an outlier in a hierarchical clustering based on pairwise Pearson correlation between all samples in PRECISE-1K; and (iv) having a minimum Pearson correlation between biological replicates of 0.95.

4.4.7 Ribosome Profiling

Ribosome profiling libraries were created using a modified version of the protocol outlined in Latif et al [237]. The protocol was modified to negate the effects of the addition of chloramphenicol by grinding frozen cells. 50 mL of cell culture was harvested by centrifugation for 4 minutes at 37°C in a 50 mL conical tube containing 0.4 g of sand. Supernatant was aspirated quickly and the pellet was flash frozen in liquid nitrogen. Pellets were transferred into a liquid nitrogen cooled mortar and pestle, 500 µL of lysis buffer was added, and the pellet was pulverized to lyse the cells. Lysate was transferred to a falcon tube to thaw on ice. The lysate was then centrifuged, and the supernatant was isolated to continue with the published protocol. Reads were sequenced on an Illumina HighSeq machine using a single end 50 bp kit.

Adaptors were removed from ribosome profiling reads using CutAdapt v1.8 [238], then mapped to the *E. coli* K-12 MG1655 reference genome (NC_000913.3) using bowtie [233]. They were scored at the 3' end to generate ribosome density profiles.

4.4.8 Generation of *pitA* Mutants

The mutations referred to in **Figures 4.5I-K and B.2** were introduced into the starting (0_0) and evolved (1_0) genomes using a Cas9-assisted Lambda Red homologous recombination method. Golden gate assembly was first used to construct a plasmid vector harboring both Cas9 and lambda red recombinase genes under the control of an L-arabinose inducible promoter, a single guide RNA sequence, and a donor fragment generated by PCR which contained the desired *pitA* +T mutation and around 200 bp flanking both sides of the Cas9 target cut site as directed by the guide RNA. After allowing cells harboring the plasmid to grow for 2 hours at 30°C, L-arabinose was added to the media and the cells were allowed to grow for 3 to 5 hours, at which time a portion of the culture was plated. Single colonies were screened using ARMS PCR. Amplicons spanning the mutation site, generated with primers annealing to the genome upstream and downstream of the sequence of the donor fragment contained in the plasmid, were confirmed with Sanger sequencing. Confirmed isolates were cured of the plasmid by growth at 37°C.

4.4.9 Cell Motility Assay

We performed motility assays in duplicate for each of the conditions shown in **Figure B.2**. We mixed a tryptone broth (13 g tryptone and 7 g NaCl per liter of media) with 0.25% agar and the desired PQ level. We autoclaved the broths, then poured 25 mL into petri dishes and solidified them at room temperature overnight. Fresh colonies were spotted in the middle of the semi-solid agar with a toothpick. The plates were then incubated at 37°C for 6–8 hours and imaged on a Gel Imaging System.

4.4.10 iModulon Computation and Curation

The full PRECISE-1K compendium, including the samples for this study, was used to compute iModulons using our previously described method [32], [140]. The log(TPM) dataset **X** was first centered such that wild-type *E. coli* MG1655 samples in M9 minimal media with glucose had mean expression values of 0 for all genes. Independent component analysis was performed using the Scikit-Learn (v0.19.0) implementation of FastICA [24]. We performed 100 iterations of the algorithm across a range of dimensionalities, and for each dimensionality we pooled and clustered the components with DBSCAN to find robust components which appeared in more than 50 of the iterations. If the dimensionality parameter is too high, ICA will begin to return single gene components; if it is too low, the components will be too dense to represent biological signals. Therefore, we selected a dimensionality which was as high as possible without creating many single gene components, as described [32]. At the optimal dimensionality, the total number of iModulons was 201. The output is composed of matrices **M** [genes x iModulons], which defines the relationship between each iModulon and each gene, and **A** [iModulons x samples], which contains the activity levels for each iModulon in each sample.

For each iModulon, a threshold must be drawn in the **M** matrix to determine which genes are members of each iModulon. These thresholds are based on the distribution of gene weights. The highest weighted genes were progressively removed until the remaining weights had a D'agostino K^2 normality below 550. Thus, the iModulon member genes are outliers from an otherwise normal distribution. iModulon annotation and curation was performed by comparing them against the known TRN from RegulonDB [11]. Names, descriptions, and statistics for each iModulon are available from the PRECISE-1K manuscript [140] and iModulonDB [21].

4.4.11 Differential iModulon Analysis

DiMAs were calculated as previously described [19], [128]. For each iModulon, a null distribution was generated by calculating the absolute difference between each pair of biological replicates and fitting a log-normal distribution to them. For the groups being compared, their mean difference for each iModulon was compared to that iModulon's null distribution to obtain a p-value. The set of p-values for all iModulons was then false discovery rate (FDR) corrected to generate q-values. Activities were considered significant if they passed an absolute difference threshold of 5 and an FDR of 0.1. The main comparison in this study was between the starting strain at 250 μM PQ (n = 2) and the combined set of all evolved strains at 250 and 750 μM PQ (n = 61). Performing the comparison using both concentrations of PQ ensures that our comparison captures all of the major effects of tolerization. The set of DiMAs was similar when performing the comparison at just one or the other concentration.

We also performed a brief DEG analysis, which used the same algorithm as above but with individual gene expression values instead of iModulon activities.

4.4.12 iModulon Explained Variance Calculation

The explained variance for each iModulon in this study was calculated using our workflow [128]. Since iModulons are built on a matrix decomposition, the contribution of each one to the overall expression dataset can be calculated. For each iModulon, the column of **M** and the row of **A** for the evolved samples in this study were multiplied together, and the explained variance between the result and the full expression dataset was computed. These explained variance scores were used to size the subsets of the treemap in **Figure 4.2H**. Note that the variance explained by ICA is 'knowledge-based' in contrast to the 'statistic-based' variance explanation provided by the commonly used principal component analysis (PCA).

4.4.13 Metabolism and Expression Modeling with OxidizeME

We used OxidizeME, a genome-scale model of metabolism and expression (ME) with ROS damage responses [126]. Models used for flux maps were constrained using phenotypic data (glucose uptake rate and growth rate) and expression data as previously described [149], [150]. In order to force

PQ cycling in the model, the lower bounds for the 'PQ2RED_FWD_FLAVONADPREDUCT-MONOMER_mod_fad' and 'PQ1OX_FWD_SPONT' were set to the same non-zero value and iterated over. Additionally, the former reaction was amended to accept NADH as an electron donor by editing the stoichiometry. PQ cycling sweeping calculations were performed by sampling various lower bounds to identify the range the model could support growth, and then sweeping 100 uniform values within that range. The total NADH produced through the TCA cycle was calculated by summing the fluxes for the 'MDH' and 'AKGDH' metabolic reactions.

Table 4.1: Complexes from the OxidizeME model used to calculate damage to the TCA cycle by oxidative stress.

ComplexFormation Reaction ID	**Associated Protein**
damage_SUCC-DEHASE_mod_3fe4s_mod_fad_mod_2fe2s_mod_4fe4s_o2s	Succinate Dehydrogenase
damage_CPLX0-7760_mod_4fe4s_o2s	Aconitase A
damage_CPLX0-7761_mod_4fe4s_o2s	Aconitase B
damage_FUMARASE-A_mod_4fe4s_o2s	Fumarase A
damage_FUMARASE-B_mod_4fe4s_o2s	Fumarase B

The percentage of the proteome allocated to the TCA cycle was calculated using the solutions from each model, specifically the translation fluxes:

$$\%\ Proteome\ Allocated\ to\ the\ TCA\ cycle\ =\ \frac{\sum_{i} mw_i * V_i^{translation}}{\sum_{j} mw_j * V_j^{translation}}$$

Where mw_i and $V_i^{translation}$ represents the molecular weight and translation flux of the *i*th protein in the TCA cycle, and mw_j and $V_j^{translation}$ represents the molecular weight and translation flux of the jth protein the entire model. The damaged portion of the proteome was calculated as follows:

$$\%\ Damaged\ Proteome\ Allocated\ to\ the\ TCA\ cycle\ =\ \frac{\sum_{k} mw_k * V_k^{complexformation}}{\sum_{i} mw_j * V_j^{translation}}$$

Where mw_j and $V_j^{translation}$ are the same variables above, and mw_k and $V_k^{complexformation}$ correspond to the kth protein in **Table 4.1**. The undamaged portion of the proteome allocated to the TCA cycle was calculated as the difference between the total proteome allocated and the damaged proteome allocated.

4.5 Acknowledgements

This work was funded by the Novo Nordisk Foundation Grant Numbers NNF10CC1016517 and NNF20CC0035580, and by National Institute of General Medical Sciences grant number GM057089. It used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We would like to thank Marc Abrams for his assistance with manuscript editing and Lachlan Munro for assistance with wet lab experiments. We would also like to thank Daniel Zielinski, Lei Yang, Emre Ozdemir, Tobias Alter, Ke Chen, and Jayanth Krishnan for helpful discussions. We thank Jason Yang for invaluable discussions and advice on an early draft of this manuscript. **Figures 4.1 and 4.3B** were created in Biorender.

Author Contributions: K.R., J.T, C.A.O., J.H.P, L.Y., A.M.F, and B.O.P. designed the study. J.T., J.H.P, R.S., Y.H., A.A., C.A.O., J.J., and E.T.T.M performed experiments. K.R., J.T., A.V.S., P.V.P., and C.L. analyzed the data. A.P. performed simulations. K.R., J.T., and B.O.P. wrote the manuscript, with contributions from all the other co-authors.

Chapter 4, in part, is currently under review for publication: Rychel K, Tan J, Patel A, Lamoureux C, Hefner Y, Szubin R, Johnsen J, Mohamed ETT, Phaneuf PV, Anand A, Olson CA, Park JH, Sastry AV, Yang L, Feist AM, Palsson BO. Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance. 2023. The dissertation author is the primary author.

Chapter 5. Laboratory evolution reveals transcriptional mechanisms underlying thermal adaptation of *Escherichia coli*

Adaptive laboratory evolution (ALE) is able to generate microbial strains with extreme phenotypes, which help reveal fundamental biological adaptation mechanisms. Here, we use ALE to evolve *Escherichia coli* strains that grow at temperatures of 45.3°C, a temperature lethal to wild type cells. The strains adopted a hypermutator phenotype that made global analysis of the DNA mutations difficult. This motivated the use of independently modulated gene set (iModulon) analysis to understand high temperature tolerance adaptation mechanisms at the transcriptomic level. Five transcriptional mechanisms underlying growth at high temperatures were revealed. These mechanisms were connected to fixed mutations, sensory inputs, and phenotypes. They are: (i) downregulation of general stress responses while upregulating the specific heat stress response; (ii) upregulation of flagellar basal bodies without upregulating motility, and upregulation fimbriae; (iii) shift toward anaerobic metabolism to avert autoxidation, (iv) regulation of iron uptake, and (v) upregulation of *yjfIJKL*, a novel heat tolerance operon. These five mechanisms explain nearly half of all variance in the gene expression of adapted strains. These thermotolerance strategies reveal that streamlining stress responses and metabolism can be achieved with a small number of simple regulatory mutations, and may suggest a new role for large protein export systems. ALE with transcriptomic characterization is a productive approach for elucidating and interpreting adaptation to otherwise lethal stresses.

5.1 Introduction

Adaptive laboratory evolution (ALE) selects for microbes that push biological systems to extremes, which enables the study of interesting new phenotypes and can provide insight into fundamental biology. Starting with a microbe and condition of interest, cells are grown for many generations and propagated in exponential growth phase when flasks reach a target density [144]. Mutants that spontaneously arise in the flask and are able to grow faster under the given condition will have a higher likelihood of being propagated, so the population will accumulate beneficial mutations and evolve. For tolerization ALEs, cells are grown in a stressful condition, and the stress is increased when the flask's growth rate stabilizes [145], enabling the development of highly stress-tolerant strains. A detailed understanding of these strains reveals mechanisms of stress tolerance which are able to inform the design of cellular factories [145], our understanding of the evolution of pathogens [147], and the fundamental science of systems that may otherwise be hard to study.

Typically, ALE endpoint strains are studied by DNA resequencing and subsequent characterization of the mutations that are expected to improve fitness. This works well for many ALEs, as the average evolved strain has only ~22 mutations (according to ALEdb, a database of such mutations [146]). However, microbes are able to increase their mutation rate by mutating the DNA mismatch repair machinery, and will evolve into hypermutator strains in highly stressful environments [239]. These strains are therefore of particular interest, since they rapidly acquire novel phenotypes. However, it is very difficult to elucidate the key causal mutations from hypermutator strains due to the high number of mutations, each of which may interact with one another in complex ways or have very little effect at all. Therefore, methods that can reduce the complexity of analyzing hypermutator strains and cut through the genomic noise of their many mutations are needed.

The transcriptional regulatory network (TRN) senses cellular states and environments, and helps to maintain homeostasis and regulate growth by adjusting gene expression levels. Analyzing changes to transcriptomic allocation in hypermutator strains represents a possible route for their characterization, as it can reveal regulatory changes induced by the mutations or be used to infer changes to metabolism and stress. However, we again run into a problem of scale: hundreds or thousands of genes may be differentially expressed, making it difficult to glean a global understanding from transcriptomic datasets.

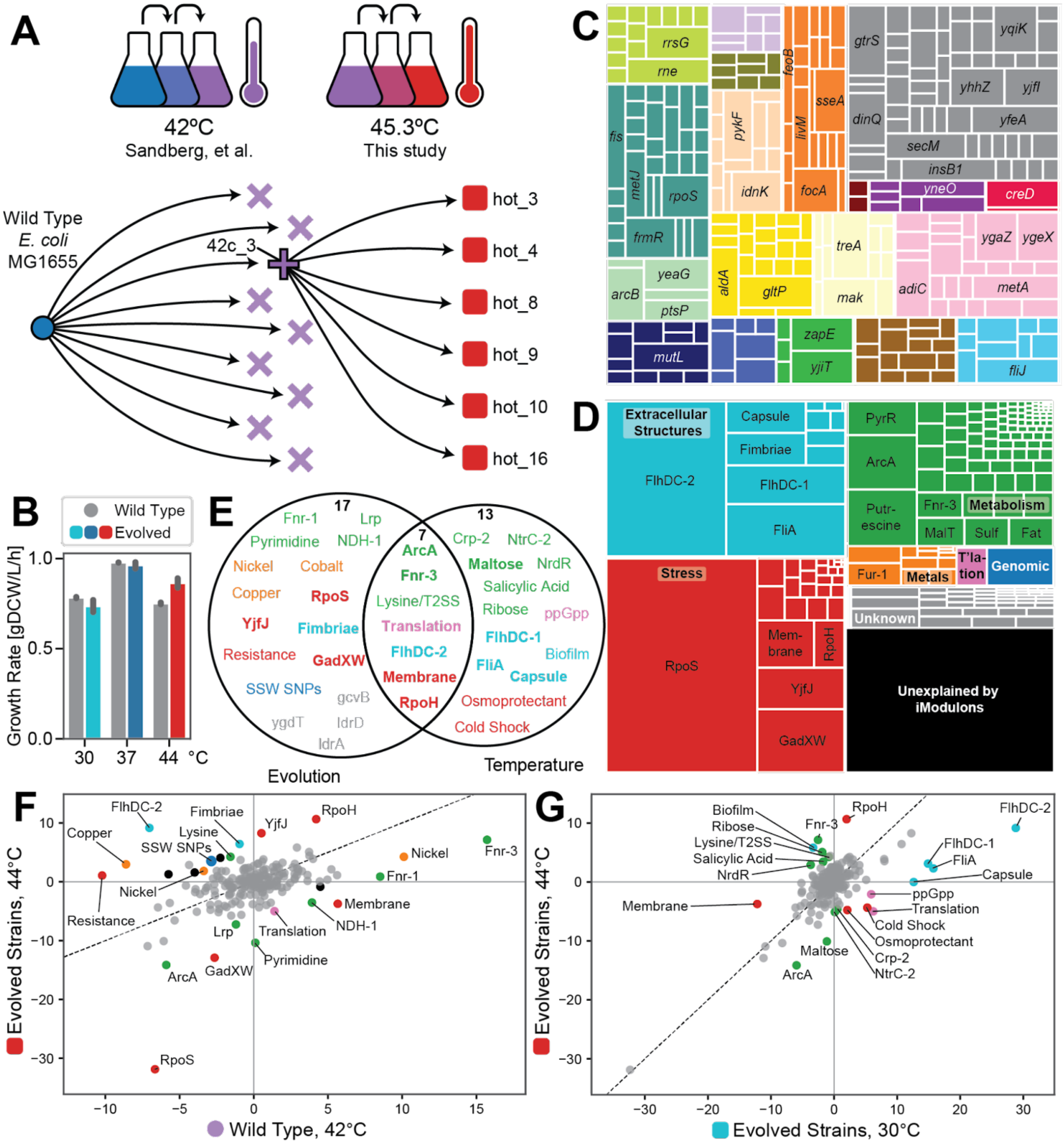
This problem can now be addressed by a recently developed approach called independently modulated gene set (iModulon) analysis [19]. iModulon analysis employs independent component analysis (ICA) to identify co-regulated signals from large compendia of gene expression data. These signals are represented by iModulons, which have a weighting for each gene and an activity level in each sample. Highly weighted genes are considered to be members of the iModulon, and they tend to match very well to regulons defined experimentally. Analyzing the activity levels of iModulons decreases the number of significant variables approximately 17-fold [140], making it tractable to characterize most of the variance in gene expression. Since they are trained on dozens or even hundreds of transcriptomes, they also provide useful context for analyzing trends in sets of samples. iModulon structures have been established for several organisms, including *E. coli* [19], [140], and are available to browse, search, or download from iModulonDB.org [21]. iModulons have proven useful for analyzing non-hypermutator ALE strains in several cases (including **Chapter 4**) [39], [40], [149]–[151], [240], making them a promising option for characterizing hypermutator strains.

To explore how transcriptomes evolve, we must choose a selection pressure that will produce informative strains. High temperature exerts a fundamental stress on biological systems by destabilizing proteins and other molecules [153]. Tolerating this stress has driven evolution since life began [241], is relevant to understanding the response of pathogens to fever [242], and could be helpful to engineer more efficient cell factories [243]. A prior study used ALE with increasing temperatures to generate ten E. coli strains that grow well at 42°C [244]. Mutational analysis and a simple transcriptomic analysis of these strains revealed some valuable thermal adaptation strategies, such as modifying mRNA degradation and peptidoglycan recycling pathways.

Here, we evolved an isolate from the 42°C evolution even further to push the limits of heat tolerance. The six endpoint strains of this study can grow at temperatures as high as 45.3°C, which is lethal to wild type strains. To achieve this increase in heat tolerance, the strains were all hypermutators.

We generated transcriptomic samples from these strains at various temperatures, included them in a compendium of over 1,000 RNA sequencing (RNAseq) datasets called PRECISE-1K [140], and extracted iModulons. We enumerated each of the major transcriptomic adaptations that facilitate rapid growth at high temperatures. Despite their broad range of genomic mutations, the strains exhibited only a few major iModulon changes in their transcriptomes. For several iModulon changes, we describe potential mechanisms based on the observed regulatory mutations and a review of the literature. We reveal changes to the regulation of stress responses, motility, redox metabolism, and iron uptake. We also propose that a previously uncharacterized operon, *yjfIJKL*, which was upregulated in these strains, is beneficial for survival at high temperatures. In addition to the specific insights on heat tolerance, this study demonstrates the value of iModulon analysis for gaining clear insights from hypermutator strains.

Figure 5.1: ALE increased heat tolerance via changes to the genome and transcriptome. **(A)** ALE schematic, showing a previous round of ALE that generated *E. coli* strain that tolerated up to 42°C [244], and that this study focuses on descendents from a single strain of the prior study to generate strains that tolerate up to 43.5°C. Symbol shapes represent strain cohorts, and colors represent temperatures; these will be kept consistent throughout the paper. See **Figure C.1** for details. **(B)** Growth rates for the wild type and final evolved strains at three temperatures, showing a significant increase in growth rate at 44°C ($p = 5.9*10^{-7}$). **(C)** Treemap of all 504 mutations observed in any of the six evolved strains, where each mutation is mapped to its nearest gene and genes that mutate in five or more strains are labeled. Colors indicate clusters of orthologous genes (COGs). **(D)** Treemap of the variance in the transcriptomes of the evolved strains by iModulon, showing that relatively few iModulons capture most of the variation. The 20 iModulons which explain the most variance are labeled, with some names shortened for space. For more information on each iModulon, see iModulonDB.org. **(E)** Venn diagram of the significantly different 37 iModulon activities (DiMAs) from evolution (F) and temperature changes (G). Colors match the categories in (D). iModulons with bolded names are also in the top 20 by explained variance. **(F-G)** Colors match the categories in (D), except that gray represents insignificant iModulon activities and black represents the "unknown" category. **(F)** DiMA plot comparing the iModulon activities in the wild type and evolved strains at their highest respective temperatures. **(G)** DiMA plot comparing iModulon activities in the evolved strains at cold (30°C) and hot (44°C) temperatures.


A
42ºC
Sandberg, et al.
45.3ºC
This study
Wild Type
E. coli
MG1655
42c_3
hot_3
hot_4
hot_8
hot_9
hot_10
hot_16
B
Wild Type
Evolved
Growth Rate [gDCW/L/h]
1.0
0.5
0.0
30
37
44
°C
C
rrsG
rne
fis
metJ
rpoS
frmR
arcB
yeaG
ptsP
mutL
pykF
idnK
aldA
gltP
feoB
livM
sseA
focA
zapE
yjiT
treA
mak
gtrS
dinQ
yqiK
yhhZ
yjfI
yfeA
secM
insB1
yneO
creD
ygaZ
ygeX
adiC
metA
fliJ
D
Extracellular Structures
Capsule
Fimbriae
FlhDC-2
FlhDC-1
FliA
PyrR
ArcA
Putr-escine
Fnr-3
Metabolism
MalT
Sulf
Fat
Fur-1
Metals
T'la-tion
Genomic
Stress
RpoS
Mem-brane
RpoH
YjfJ
GadXW
Unknown
Unexplained by iModulons
E
17
7
13
Fnr-1
Lrp
Pyrimidine
NDH-1
Nickel
Cobalt
Copper
RpoS
YjfJ
Fimbriae
Resistance
GadXW
SSW SNPs
gcvB
ygdT
ldrD
ldrA
ArcA
Fnr-3
Lysine/T2SS
Translation
FlhDC-2
Membrane
RpoH
Crp-2
NtrC-2
Maltose
NrdR
Salicylic Acid
Ribose
ppGpp
FlhDC-1
Biofilm
FliA
Capsule
Osmoprotectant
Cold Shock
Evolution
Temperature
F
FlhDC-2
Fimbriae
RpoH
YjfJ
Copper
Lysine
SSW SNPs
Nickel
Fnr-3
Fnr-1
Nickel
Resistance
Membrane
NDH-1
Translation
Lrp
GadXW
Pyrimidine
ArcA
RpoS
Evolved Strains, 44°C
Wild Type, 42°C
10
0
−10
−20
−30
−10
−5
0
5
10
15
G
Biofilm
Ribose
Fnr-3
RpoH
FlhDC-2
Lysine/T2SS
Salicylic Acid
NrdR
FlhDC-1
FliA
Capsule
Membrane
ppGpp
Translation
Cold Shock
Osmoprotectant
Maltose
Crp-2
NtrC-2
ArcA
Evolved Strains, 44°C
Evolved Strains, 30°C
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

5.2 Results

5.2.1 ALE increased heat tolerance via a hypermutator phenotype

Using ALE, we obtained six evolved strains that tolerated 45.3°C (**Figure 5.1A, C.1**). Each one descended from the same ancestor from the previous 42°C ALE [244], 42c_3, which itself descended from *E. coli* K-12 MG1655. We generated growth curves and computed growth rates for each of the strains at 30, 37, and 44°C (**Figure 5.1B**), showing a significant increase in growth rate at 44°C after evolution ($p = 5.9*10^{-7}$). Interestingly, the strains did not exhibit a major tradeoff in growth rates at 37°C on average ($p = 0.13$), and only had a slight growth disadvantage against wild type at 30°C ($p = 0.012$). The evolved strains maintained much of their ability to adapt to changes in temperature, suggesting that the genomic mutations were not especially deleterious.

The ancestral strain, 42c_3, contained 30 mutations, including *mutL* G49V. MutL is part of the DNA mismatch repair machinery, and mutations in this gene tend to induce increases in mutation rates [239], [245]–[247]. Thus, each of the evolved strains was a hypermutator, and ended with between 60 and 126 mutations, with the average strain experiencing 84 mutations. Each of the mutations was assigned to its nearest gene, and mutated genes were visualized in **Figure 5.1C**. No particular cluster of orthologous genes (COG) was enriched in this set, and the large number of mutations precluded a detailed analysis of the potential benefit of each one.

5.2.2 iModulon analysis revealed a small set of transcriptomic adaptations

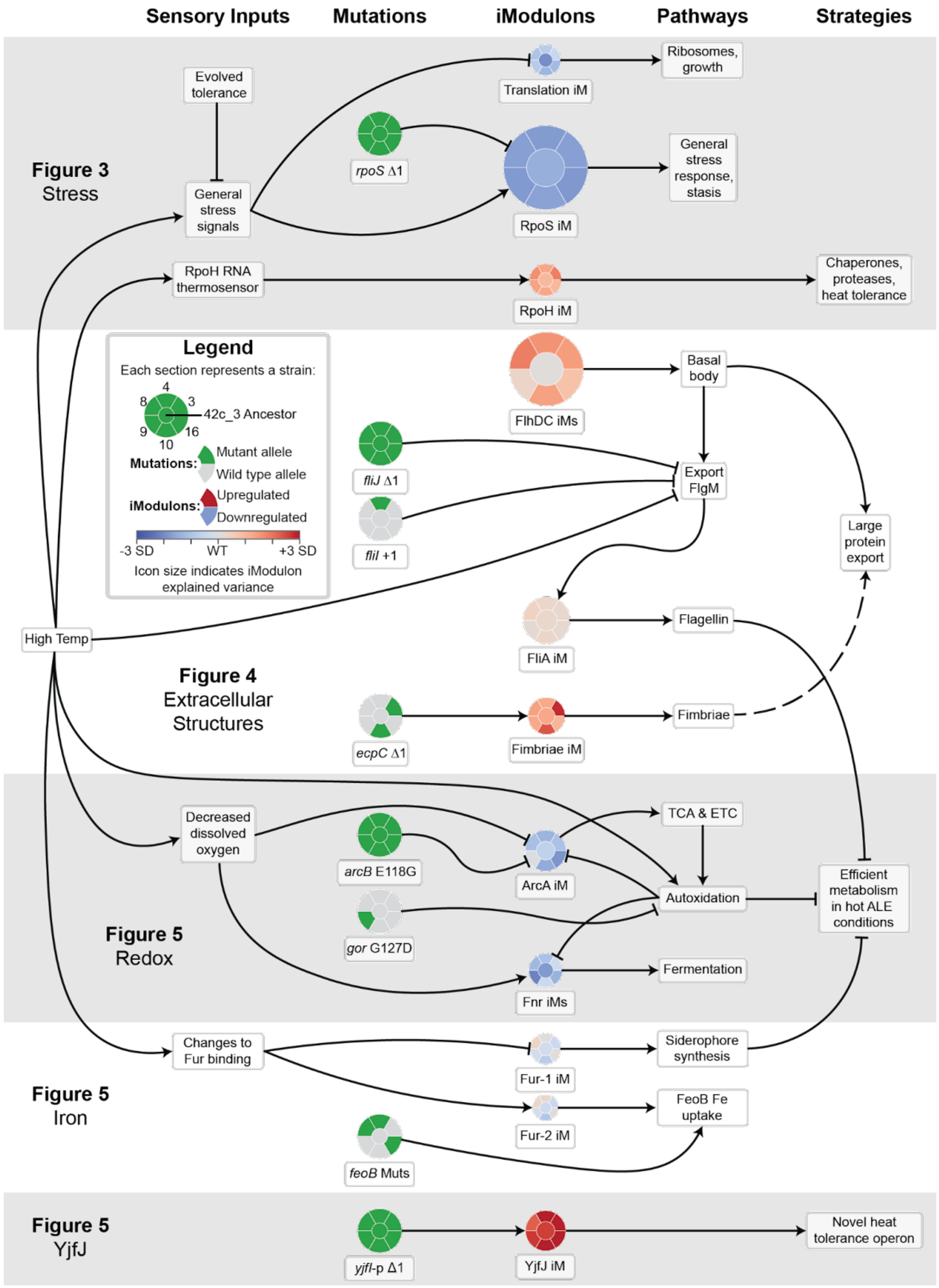
In order to gain a clear understanding of the adaptations in the evolved strains, we generated RNAseq data for each of them and the 42c_3 ancestor at 30, 37, and 44°C in duplicate. From the prior study [244], we also had samples at 42°C for the wild type and each of the 42°C evolved strains. All of these profiles were included in PRECISE-1K, a compendium of over 1,000 *E. coli* transcriptomes which were generated using the same experimental protocol and analyzed with iModulons in aggregate [140].

PRECISE-1K provides a large and diverse condition space for ICA to identify co-regulated, independently modulated signals (iModulons). The 201 iModulons computed from PRECISE-1K have been characterized with assigned functions, regulators, and categories to facilitate interpretation. They are available at iModulonDB.org [21] under "*E. coli* PRECISE-1K", in the project "hot_tale".

Because ICA is a matrix decomposition method [22], any subset of iModulons can be used to infer the original gene expression data across any subset of conditions in PRECISE-1K. Therefore, we can quantify the explained variance of each iModulon in the samples from the evolved strains (**Figure 5.1D**). In the data generated for this study (a subset of PRECISE-1K), the 201 iModulons captured 82% of the variance in the data. The top 20 highest explained variance iModulons explained 61.3% of the overall variance. Thus, a relatively small number of variables can be highly informative about the global state of these samples, and therefore represent an approach to identify the key thermotolerance strategies that emerge during ALE.

Differential iModulon activities (DiMAs) are similar to differentially expressed genes (DEGs), except they are much easier to interpret because the iModulons are much fewer in number (201) than genes (4257). iModulons are also knowledge-enriched with regulatory information [140]. DiMAs between starting and evolved strains represent a summary of transcriptional adaptations [240] (**Figure 5.1F**). In this dataset, we can also quantify DiMAs for the evolved strains between the cold and hot temperatures (**Figure 5.1G**), and compare the sets of significant iModulons (**Figure 5.1E**).

Figure 5.2: Overview of the five adaptive mechanisms to high temperature growth. On the left side, high temperature is the main "input", which leads to a variety of sensory inputs in the first column, which are expected from literature or inferred from the iModulon evidence. In the second column, mutations in the evolved strains are shown, according to the legend. iModulons in the third column integrate sensory inputs and effects of mutations to determine their activity levels. Colors in the iModulon icons represent the change between the wild type (WT) activity at 42°C and the given strain's evolved activity at 44°C, normalized by the standard deviation (SD) of the iModulon's activity in all samples from PRECISE-1K. iModulon icons are sized according to their explained variance (from Fur-2, 0.27% to RpoS, 19.8%; scaled using the square root). To the right of the iModulon column are the pathways and phenotypes which are determined by the transcriptomic and genomic changes. The far right column lists hypothesized strategies by which the evolved strains tolerated heat. Different background shades represent different topics, and each is labeled with the respective figure from this chapter in which to find more information. Dotted lines represent lower confidence relationships that warrant further research.


Sensory Inputs
Mutations
iModulons
Pathways
Strategies
Figure 3
Stress
Evolved tolerance
General stress signals
rpoS Δ1
Translation iM
Ribosomes, growth
RpoS iM
General stress response, stasis
RpoH RNA thermosensor
RpoH iM
Chaperones, proteases, heat tolerance
Legend
Each section represents a strain:
4
3
8
42c_3 Ancestor
9
16
10
Mutations:
Mutant allele
Wild type allele
iModulons:
Upregulated
Downregulated
-3 SD
WT
+3 SD
Icon size indicates iModulon explained variance
FlhDC iMs
Basal body
fliJ Δ1
fliI +1
Export FlgM
Large protein export
High Temp
FliA iM
Flagellin
Figure 4
Extracellular Structures
ecpC Δ1
Fimbriae iM
Fimbriae
Decreased dissolved oxygen
arcB E118G
ArcA iM
TCA & ETC
gor G127D
Autoxidation
Efficient metabolism in hot ALE conditions
Figure 5
Redox
Fnr iMs
Fermentation
Changes to Fur binding
Fur-1 iM
Siderophore synthesis
Figure 5
Iron
Fur-2 iM
FeoB Fe uptake
feoB Muts
Figure 5
YjfJ
yjfI-p Δ1
YjfJ iM
Novel heat tolerance operon

Highly variable and differentially activated iModulons in ALE studies typically indicate one of three features [240]: (1) large genomic alterations have directly amplified or deleted the genes of an iModulon, (2) mutations in regulatory pathways have altered gene expression, or (3) underlying metabolites or processes which are sensed by the TRN have been altered. There were no major amplifications or deletions in the evolved strains which resulted in their own iModulons. Therefore, the major signals are either the result of regulatory mutations or the output of the sensory systems within the evolved cells. Based on the direction of the change, the large body of existing literature on the *E. coli* TRN, and experimental evidence, we have inferred the mechanisms which underlie the major changes in the transcriptome. We also proposed explanations of how they provide benefits to the evolving strains. We present the five mechanisms with the strongest signals in the following sections, and in **Figure 5.2**.

5.2.3 Stress sigma factors shift allocation from general to specific responses

The iModulon with the largest explained variance is the RpoS iModulon, which reallocates an enormous 19.8% of the transcriptome in the evolved strains. RpoS is the general stress response sigma factor, which is governed by complex regulation and limits growth when active [190], [191]. Prior iModulon studies have explored a "fear-greed tradeoff", where typical strains exhibit a negative correlation between the RpoS and Translation iModulon activities. Faster growing cells activate the Translation and downregulate the RpoS iModulons (see **Section 4.2.8.1**) [19], [40], [43], [192], [193]. As the prior generation of 42°C evolved strains mutated to tolerate high temperatures, they experienced less stress and therefore downregulated RpoS [244] (**Figure 5.3A,** "42c Other"). In the 42c_3 ancestor of the high temperature tolerant strains, we observe even stronger downregulation of the RpoS iModulon, which is maintained in the new 45°C evolved strains.

This iModulon activity is likely resulting from a frameshift mutation in *rpoS* in 42c_3 and its derivatives. The mutation appears to have mostly deactivated RpoS, allowing sigma factors that do not suppress growth to outcompete it and providing a growth rate benefit during ALE. Deactivating RpoS is generally a good strategy for ALE [19], [40], [43], [192], [193], but the temperature adapted strains take this to an extreme via this mutation.

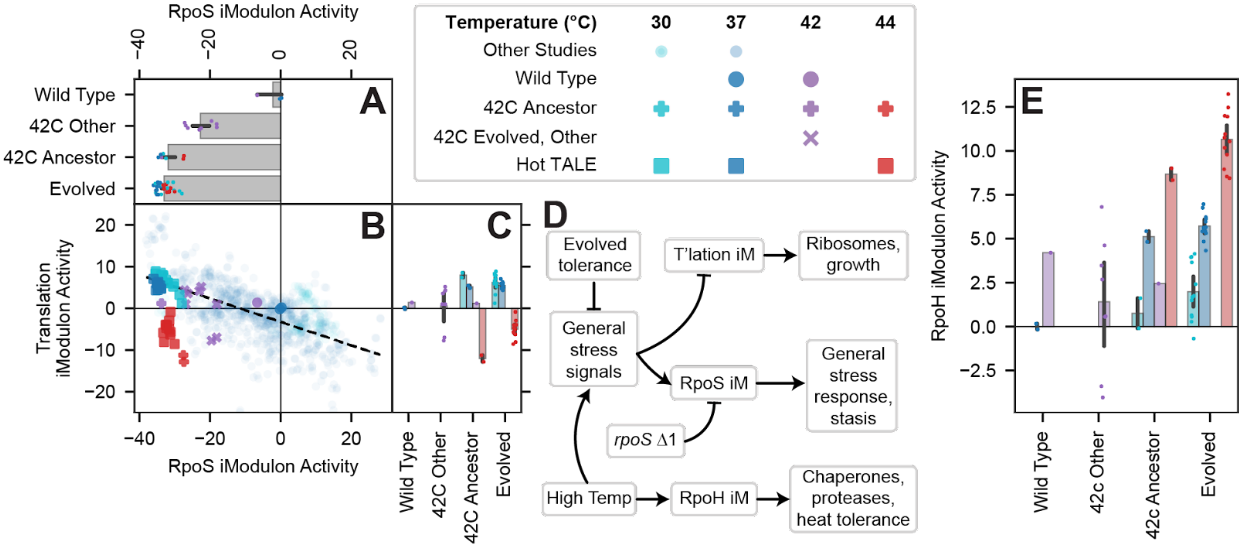


Figure 5.3: Stress sigma factors shift allocation from general to specific responses. Error bars represent mean ± 95% confidence interval. Colors in the columns of the legend are consistent for each plot. **(A-C)** Activities of the RpoS and Translation iModulons, which constitute the fear-greed tradeoff. **(A)** RpoS activity is downregulated by evolution (p = 0.036), and accounts for 19.8% of the variance in the dataset. This iModulon regulates the general stress response which slows growth. Since very little variation by temperature was observed, bars represent strains regardless of temperature. **(B)** Scatter plot labeled according to the legend, with low opacity circles representing all other samples in PRECISE-1K. A black dashed line was fit to the other samples, representing the typical fear-greed tradeoff. The evolved samples have lower RpoS activity than expected, due to the *rpoS* mutation in the strains. **(C)** Translation activity is correlated with temperature, but downregulated less strongly after evolution due to successful adaptation (p = 0.027). **(D)** Knowledge graph explaining the three iModulons in this figure. **(E)** RpoH iModulon activity, which maintains its correlation with temperature but is slightly upregulated by prolonged heat exposure (p = 0.0031). RpoH regulates high temperature responses.

The Translation iModulon is typically anti-correlated with RpoS, because similar underlying growth/stress and RNA polymerase-related variables control both iModulons [190], [191]. However, the Translation iModulon remains anti-correlated with temperature (**Figure 5.3C**) while the RpoS iModulon is downregulated at all temperatures, diverging from the usual fear-greed tradeoff (**Figure 5.3B**). The *rpoS* frameshift stops RpoS activity but does not regulate the Translation iModulon, explaining this discrepancy (**Figure 5.3D**). We do observe a small upregulation of the Translation iModulon at high temperatures after evolution, suggesting that the mutations and tolerization strategies in these strains have successfully decreased the stress signals which typically downregulate Translation at high temperatures.

Unlike RpoS, the heat stress sigma factor RpoH does not mutate, maintains its wild type correlation with temperature, and is differentially upregulated at high temperatures after evolution (**Figure 5.3D**). RpoH senses temperature via several mechanisms including an RNA-thermosensor and temperature-dependent proteolysis [248], [249], and it activates a variety of heat shock genes and chaperones [250]. Presumably, any mutations or changes to heat stress regulation were selected against. Prolonged exposure to high temperatures slightly upregulates RpoH, probably via the known temperature-dependent pathways.

Thus, the evolved cells downregulate general stress responses (RpoS) to improve growth, but upregulate specific responses to heat (RpoH). This represents an effective strategy for stress tolerization ALE; indeed, it mirrors the response of oxidative stress evolved strains, which maintain activity of the specific oxidative stress response, SoxS, (**Section 4.2.6.4**) while also downregulating RpoS (**Section 4.2.8.1**) [240].

5.2.4 Motility iModulons amplify basal body expression while suppressing FliA

A major fraction (18%) of the variance in the transcriptome of the thermotolerant strains is explained by motility iModulons, which respond to two transcription factors, FlhDC and FliA. The regulation of this system has been studied in detail [251], and the iModulon gene structure matches well with the known literature. The two primary iModulons of interest are FlhDC-2 and FliA (**Figure 5.4A**). The promoter of *flhDC* integrates many signals that affect motility, and then expression of FlhDC induces flagellar synthesis in steps: first, the basal body is synthesized, and then the hook, junction, flagellin, motor, and control mechanisms are added. The timing of these steps is ensured by using a second regulator, the sigma factor FliA, which is induced by FlhDC (as part of the class I genes, purple in **Figure 5.4A**), co-regulates the intermediate steps (class II genes, orange), and solely regulates the final steps (class III genes, green) [251]. A good understanding of the regulation and dynamics of this system is important for fundamental biology, understanding host-pathogen interactions [252], and developing a toolkit for designing protein secretion systems [253], [254].

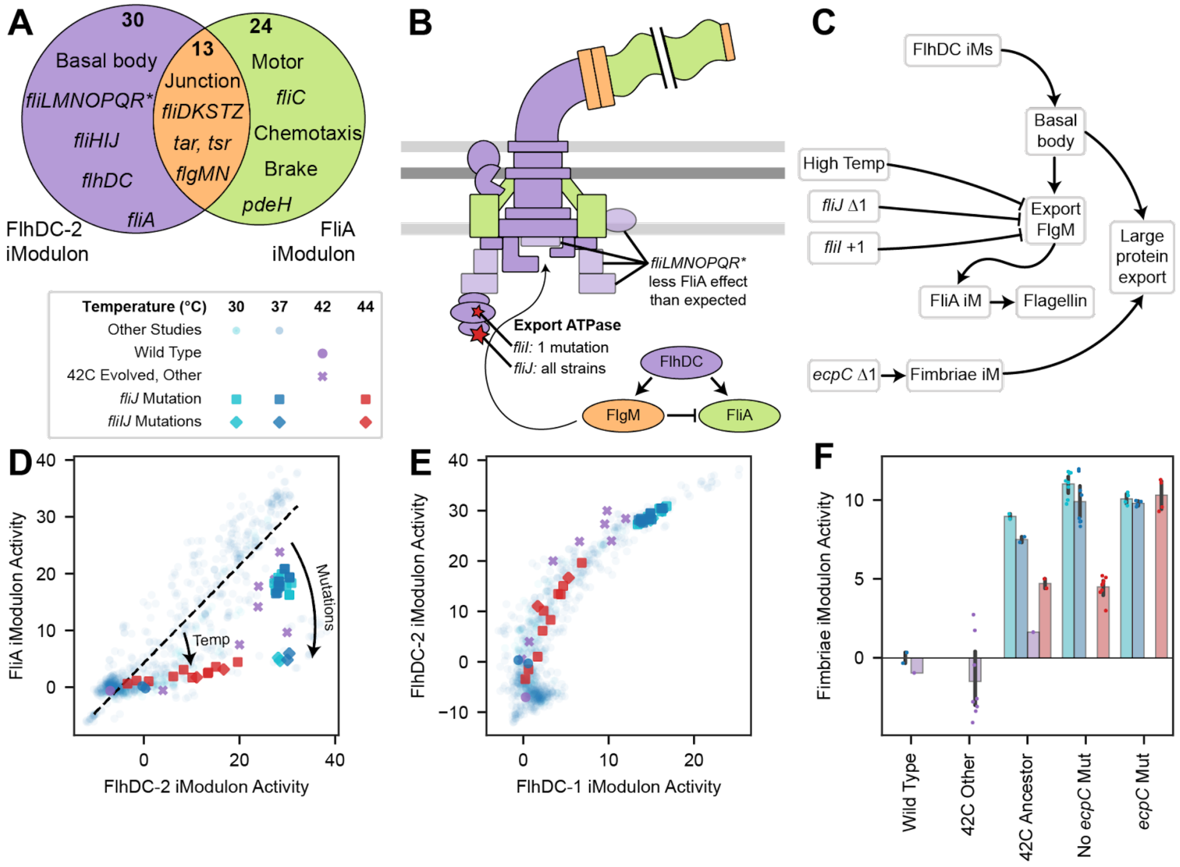


Figure 5.4: Changes to motility and fimbriae regulation suggest a possible role in protein aggregate export. **(A)** Venn diagram of genes in the two main motility iModulons, FlhDC-2 and FliA, which are highly similar to the known regulons [11], [251]. The most notable exception is *fliMNOPQR*, which is thought to be regulated by both FlhDC and FliA but is only found in the FlhDC related iModulon. **(B)** Illustration of the flagellum, adapted from [251]. Components are colored according to the Venn diagram in (A), and low opacity components are those which were expected to be under dual regulation. ATPase mutations observed in the evolved strains are pictured with a red star and label. The established regulatory cascade [251] is also pictured in the lower right, showing that the anti-sigma factor FlgM requires the ATPase to be exported and derepresses FliA. **(C)** Knowledge graph summarizing this section. **(D-E)** Scatter plots of iModulon activities (iModulon activity phase planes), illustrated according to the legend above panel (D). **(D)** Though FliA activity is typically correlated with FlhDC-2 activity (black dotted line: best fit for other projects; Pearson R = 0.87) and follows it in the regulatory cascade (B), the mutant strains (squares) do not activate FliA as strongly as expected, particularly in the case of the strain with two ATPase mutations (diamonds). **(E)** The two FlhDC iModulons form a tight curve with heat tolerant strains having high activity below 40°C but decreasing at high temperatures. **(F)** Bar and swarm plot of Fimbriae iModulon activity, with all evolved strains upregulating it ($p = 0.00030$) and those with the *ecpC* Δ1 mutation upregulating it the strongest at high temperatures. Error bars represent mean ± 95% confidence interval.

The two iModulons nearly perfectly mirror the known regulation (**Figure 5.4B**). However, FlhDC-2, but not FliA, includes *fliLMNOPQR*, despite the fact that it has binding sites for both regulators. This operon is needed earlier in the synthesis of flagella [255]. iModulons learn from relationships in expression data, so the exclusion of this operon from FliA indicates that these genes tend to be more correlated with the early stages of synthesis, in agreement with their function and despite the ability to be regulated by the late stage regulator. In addition, a third iModulon plays a unique role: the FlhDC-1 iModulon contains several genes from both iModulons at both positive and negative weights, and appears to capture a third dimension or nonlinearity in the motility transcriptome. This may reflect different binding affinities and changing ratios between the two regulators. FlhDC-1 and 2 form a nonlinear curve, suggesting that samples adjust towards FlhDC-1 as FlhDC expression increases (**Figure 5.4E**). Thus, iModulons reflect the known transcriptional regulation while providing additional nuance which is useful for a practical understanding of the system.

At high temperatures, the flagellar secretion system is less able to secrete FlgM, the anti-sigma factor for FliA [256]. This mechanism may have evolved to help *E. coli* avoid flagellin-mediated detection by the host immune system during a fever [252]. Thus, the flagella synthesis pathway is cut off at a regulatory point between the two iModulons (**Figure 5.4C**). We observe this mechanism clearly in the activity of the evolved strains at 44°C, which have some FlhDC-2 activity but no FliA activation (**Figure 5.4D**). The activity phase plane between FlhDC-2 and FliA is thus highly informative: typically, the two iModulons exhibit a strong correlation (Pearson R = 0.87), but in cases below the best fit line, a mechanism such as the failed secretion of FlgM inhibits FliA activity.

Interestingly, we also observed the evolved samples exhibiting activity below the best fit line at lower temperatures, when heat should not be disrupting FlgM secretion (**Figure 5.4D**). This observation indicates that something besides temperature in these strains downregulates the FliA iModulon. Indeed, the ancestor (42c_3) of all evolved strains had a frameshift mutation in *fliJ*, which is a chaperone that prevents aggregation of the flagellar export substrates [255]. Since FlgM likely can’t be exported as efficiently due to this mutation, the strains occupy a unique location in the phase plane. Another mutation, a frameshift in the export ATPase *fliI*, affected only the hot_4 strain. With this mutation, FliA activity decreases even further (diamonds, **Figure 5.4D**).

iModulons have thus quantitatively captured the complex transcriptional regulation of motility and revealed the effects of temperature and mutations on the regulation of FliA. Key questions remain to be elucidated: (i) the strains strongly upregulate FlhDC, but due to its complex upstream regulation it is difficult to deduce the molecular mechanism, (ii) since FlhDC activity also decreases at high temperatures, some unknown mechanism downstream of FliA may be feeding back to regulate these genes, and (iii) there may be an evolutionary benefit to expressing FlhDC-regulated genes but not FliA-regulated genes at high temperatures.

We can speculate that the evolutionary benefit in question (iii) could arise from secretion of other temperature-sensitive proteins. Given that the flagellar basal body is able to rapidly secrete large proteins [253] and its typical secretion substrate, *fliC*, is downregulated by the evolved *fliIJ* mutants, perhaps it has been repurposed to help eliminate protein aggregates in the cells.

5.2.5 The Fimbriae iModulon is a second upregulated large protein export system

Another extracellular structure, the fimbriae, is an important part of transcriptome reallocation in the evolved strains (1.6% explained variance). This iModulon contains the fimbriae synthesis genes *fimAICDFGH* [257], which are strongly upregulated in 42c_3 and the evolved strains and negatively correlated with temperature (**Figure 5.4F).** Interestingly, the negative correlation with temperature was abolished and fimbriae were strongly upregulated at high temperatures in two evolved strains, hot_3 and hot_10, which both shared a frameshift mutation in the *ecpC* gene. EcpC is a putative usher protein for another extracellular structure, the common pilus [258]. This result suggests cross-talk between various extracellular fiber systems in *E. coli*, and it also suggests that upregulation of fimbriae may be beneficial at high temperatures.

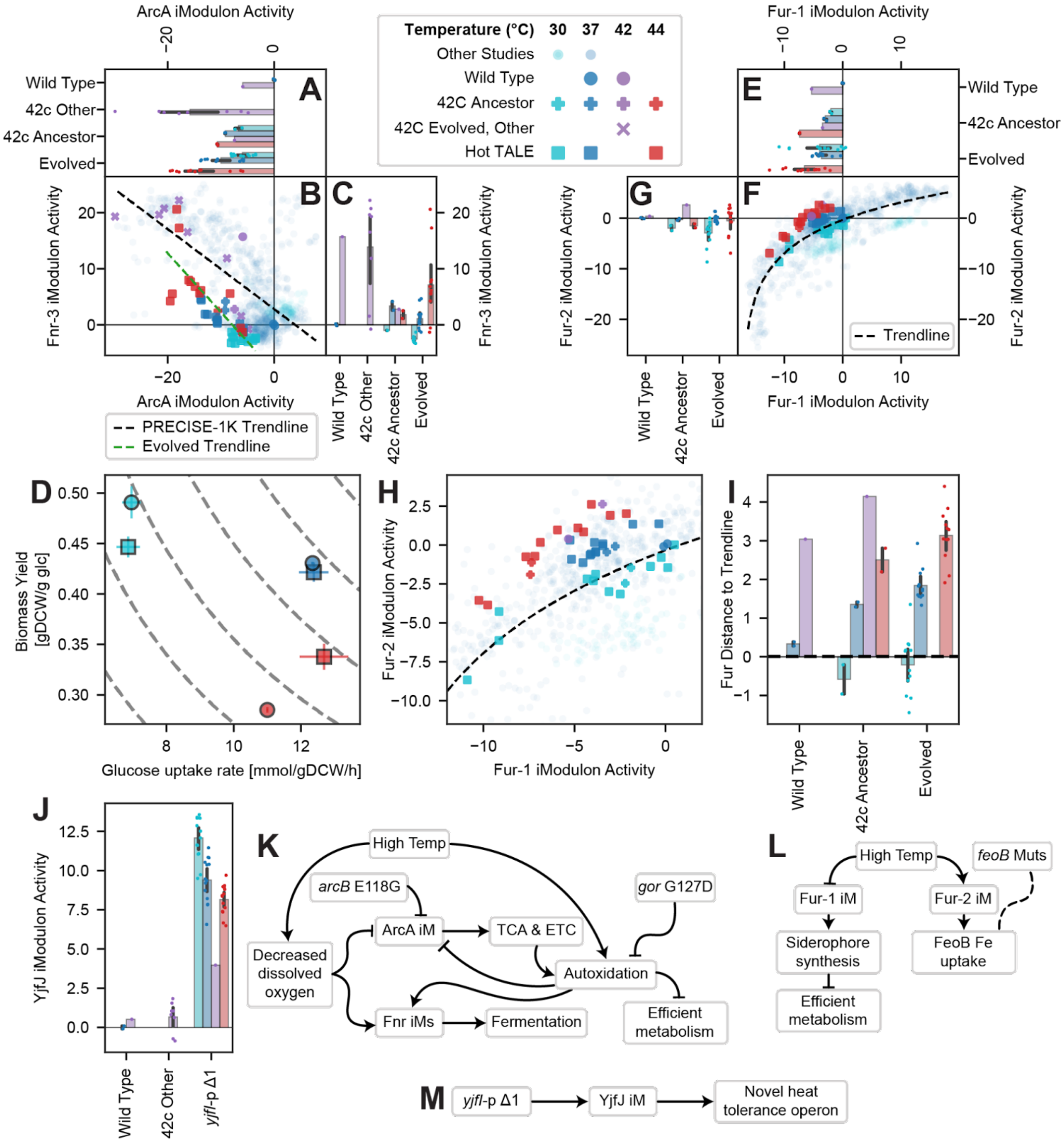
Similar to our hypothesis that misfolded proteins are exported by flagellar basal bodies, it is reasonable to also associate the expression of usher proteins *fimD* and *ecpC* with the export of misfolded proteins. In each case, an upregulated or modified gene functions to export proteins. Unlike flagellar basal bodies, the fimbriae systems are only in the outer membrane and would therefore be limited to secreting periplasmic proteins. However, they could still potentially help export misfolded proteins to support growth at high temperatures. Further research into the specificity of protein export by fimbriae and pilus systems could be helpful for the design of heterologous protein producers [259].

5.2.6 Redox Metabolism shifts toward fermentation to increase biomass yield

ArcA and Fnr, the regulators of aerobicity, exert significant control over cellular phenotypes via alterations to the expression of genes involved in respiration [218]. Together, their associated iModulons explain 3.4% of the variance in the transcriptomes of the evolved strains, but are likely to have a larger effect on metabolism and phenotypes (e.g. **Section 4.2.11.1**). The ArcAB two-component system represses aerobic metabolism genes when the electron transport chain (ETC) is in a reduced state [219], and Fnr derepresses anaerobic metabolism genes when its iron-sulfur (Fe-S) clusters are not oxidized [220]. Fnr activity is captured by three iModulons with similar activities in these samples; we therefore focus on Fnr-3, which has the highest explained variance of the three. ArcA and Fnr-3 activities are correlated (black line, **Figure 5.5B**) since they both sense different features of the same underlying cellular redox state.

As temperature increases, gene expression shifts upward and leftward in the ArcA/Fnr-3 phase plane (**Figure 5A-C**). This shift indicates that high temperatures decrease oxidation, which is consistent with the decrease in oxygen solubility as temperatures increase [260]. Decreased oxygen solubility may cause the ETC to be more reduced and the Fe-S clusters to be less oxidized, causing changes to the activity state of these iModulons. The expression change will induce a lower aero-type and tend to decrease the production of NADH and reliance on oxygen.

Figure 5.5: iModulon activities and mutations reveal hallmarks of redox metabolism, iron uptake, and uncharacterized genes which may facilitate temperature adaptation. All figures use the colors, and all scatterplots use the shapes, given in the legend (top middle). For bar graphs, error bars represent mean ± 95% confidence interval. **(A-C)** iModulon activities for ArcA, which regulates aerobic metabolism, and Fnr-3, which regulates anaerobic metabolism. In (B), linear fits for the evolved samples (green) and all other samples (black) are shown. Temperature shifts expression up the trendline toward a less oxidized state, and the evolved strains have shifted their trend leftward, likely due to an *arcB* E118G mutation. **(D)** Rate-yield plot demonstrating the effect of temperature and evolution on biomass yield and glucose uptake rate. Gray dotted lines indicate isoclines with constant growth rate. **(E-G**) iModulon activities for Fur-1 and Fur-2, which regulate iron uptake and are fit to a logarithmic curve. **(H)** Zoomed in version of (F), showing that raising the temperature tends to shift samples above the trendline, toward Fur-2 expression. Fur-2 contains *feoABC*, the simple iron transporter, whereas Fur-1 contains the more metabolically expensive and less necessary siderophore synthesis pathways. **(I)** Distance to the trendline from panels (F) and (H), showing that increasing temperatures shifts the preference of Fur toward activating Fur-2. **(J)** YjfJ iModulon activity, describing the expression of the *yjfIJKL* operon. The iModulon appears to be activated in all evolved strains by a single nucleotide promoter deletion upstream of *yjfI*. This iModulon represents an unknown molecular process, but is a clear signal detected by ICA. **(K-M)** Knowledge graphs for each of the iModulon findings presented in this figure.



Temperature (°C) 30 37 42 44
Other Studies
Wild Type
42C Ancestor
42C Evolved, Other
Hot TALE
A
ArcA iModulon Activity
−20
0
Wild Type
42c Other
42c Ancestor
Evolved
B
C
Fnr-3 iModulon Activity
20
10
0
ArcA iModulon Activity
PRECISE-1K Trendline
Evolved Trendline
Wild Type
42c Other
42c Ancestor
Evolved
E
Fur-1 iModulon Activity
−10
0
10
Wild Type
42c Ancestor
Evolved
G
F
Fur-2 iModulon Activity
−10
−20
Trendline
Wild Type
42c Ancestor
Evolved
D
Biomass Yield [gDCW/g glc]
0.50
0.45
0.40
0.35
0.30
8
10
12
Glucose uptake rate [mmol/gDCW/h]
H
Fur-2 iModulon Activity
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
Fur-1 iModulon Activity
I
Fur Distance to Trendline
4
3
2
1
0
−1
Wild Type
42c Ancestor
Evolved
J
YjfJ iModulon Activity
12.5
10.0
7.5
5.0
2.5
0.0
Wild Type
42c Other
yjfI-p Δ1
K
High Temp
arcB E118G
ArcA iM
TCA & ETC
gor G127D
Decreased dissolved oxygen
Autoxidation
Fnr iMs
Fermentation
Efficient metabolism
L
High Temp
feoB Muts
Fur-1 iM
Fur-2 iM
Siderophore synthesis
FeoB Fe uptake
Efficient metabolism
M
yjfI-p Δ1
YjfJ iM
Novel heat tolerance operon

In addition to its effect on oxygen solubility, high temperature also increases the rate of reactive oxygen species (ROS) generation by ETC components like NADH dehydrogenase (see **Section 4.2.9**) [261], [262]. Decreasing ArcA expression should decrease ETC activity and help to decrease the amount of electrons that end up being wasted by this process at high temperatures [261], [262]. Interestingly, the 42c_3 strain and all fully evolved strains harbored the mutation *arcB* E118G, suggesting that they modified the ArcAB system to better tolerate high temperatures. In **Figure 5.5B**, we observe that ArcA iModulon activity has shifted to the left of the trendline formed by the other samples [140], so we infer that this mutation increases the phosphorylation of ArcA by ArcB [263]. The *arcB* mutation would explain the shift in ArcA iModulon activities, and provide the benefit of decreasing autoxidation from the ETC, reinforcing a change induced by high temperatures. However, there ought to be a tradeoff to this mutation at lower temperatures, when autoxidation does not have as strong of an effect on biomass yield.

We also note one outlier strain, hot_9, which did not upregulate Fnr iModulons or further downregulate the ArcA iModulon when temperature increased. This strain harbored the mutation *gor* G127D, which may have enhanced ROS detoxification by glutathione reductase or decreased autoxidation [264] at high temperatures.

To explore the systems-level changes to energy metabolism that arise from these genomic and transcriptomic changes, we measured the glucose uptake rate and biomass yield of the wild type and evolved strains at three temperatures (**Figure 5.5D**). These two parameters change notably with temperature. Regions of the rate-yield plane are associated with distinct states of energy metabolism called aero-types, as has been characterized in prior studies (see **Section 4.2.10.1**) [150], [212], [240]. Samples with high biomass yields are in the highest aero-type, corresponding to efficient aerobic growth. Lower aero-types are progressively less efficient and pump fewer protons across the inner membrane during respiration. Temperature strongly affects yield and uptake: cold samples are highly efficient, but unable to rapidly uptake glucose, whereas hot samples can rapidly take up glucose but have low yield due to heat-induced damage and waste.

The evolved strains at high temperature have higher uptake and yield compared to the wild type (**Figure 5.5D**). The increased uptake may be due to the changes toward anaerobic metabolism, which utilizes more glucose [218]. The increased yield is the result of the combined success of many of the mutations which decrease temperature stress, including the decreased autoxidation brought about by the shift toward anaerobic metabolism. We note that, on average, the effects of evolution at 37°C are negligible. At 30°C, on the other hand, yield decreases while glucose uptake rates remain low. This is consistent with the *arcB* mutation preventing upregulation of the high-yield aerobic pathways, which are highly efficient in the wild type at low temperatures.

Thus, iModulon and aero-type analysis have revealed the effects of temperature and mutations on energy metabolism (**Figure 5.5K**). At high temperatures, dissolved oxygen decreases and electrons leak from the electron transport chain into ROS more readily [262], inducing a metabolic shift toward anaerobiosis which is amplified by an *arcB* mutation in the heat-tolerant evolved strains. A mutation in *gor* may alleviate some autoxidation at high temperatures. This shift successfully increases both glucose uptake and biomass yield at high temperatures, but carries a tradeoff that decreases yield at lower temperatures. This temperature tolerance strategy is informative for the fundamental biology of cross-stress tolerance and the relationship between stress and metabolism. Its mutations may also provide design variables of interest for fermentation applications which may experience high temperatures or uneven oxygenation.

5.2.7 Fur preferentially derepresses *feoB*, a commonly mutated iron transporter

The two Fur iModulons regulate iron uptake systems and exhibit a nonlinear relationship (**Figure 5.5E-G**) [43], [240]. They explain approximately 1% of the variance in the transcriptome and exhibit an interesting relationship with temperature, in which temperature shifts activity perpendicularly to the trendline (**Figure 5.5H-I**; similar to **Section 4.2.6.2**). This behavior is observed in both wild type and evolved strains, suggesting that it may be a fundamental feature of fur binding. The effect is to prefer Fur-1 in lower temperatures and Fur-2 in higher temperatures. Fur-1 contains siderophore synthesis genes (**Figure B.1A**), which are needed when iron becomes less soluble at cold temperatures, as has been studied in *Vibrio salmonicida* [265]. Fur-2, on the other hand, contains less metabolically expensive ionic iron transporters, like *feoABC* [266]. These would be preferred at higher temperatures due to their lower cost and the readily available dissolved iron (**Figure 5.5L**).

Though there are no transcriptional regulatory mutations to the iron uptake system, the transporter gene *feoB* mutates in three of the six evolved strains (hot_4: F363L; hot_8: W699*; hot_9: F363L & V563M). Further research ought to probe the effects of these mutations on the temperature stability and function of FeoABC.

5.2.8 The yjfJ operon may be a new heat tolerance operon

Finally, a large 2.2% of the explained variance in the transcriptome is attributed to a single operon of all uncharacterized genes, *yjfIJKL*, which constitutes the YjfJ iModulon. The iModulon was named as such because *yjfJ* encodes a putative transcription factor which is presumed to be the regulator of the operon. YjfJ exhibits homology to PspA, which protects against membrane stress [267]. The iModulon is upregulated in 42c_3 and the evolved strains (**Figure 5.5J**), which may have been an adaptation that supported the membrane against heat stress and therefore provided a benefit to the cells. It is likely that the upregulation was induced by a single nucleotide deletion 80 base pairs upstream of the operon (*yjfI*-pΔ1; **Figure 5.5M**). Thus, we find a putative function for these genes in heat stress whose mechanistic basis needs a detailed study to reveal the underlying molecular functions.

5.3 Discussion

Here, we used ALE to produce six *E. coli* strains which can grow at 45.3°C, a temperature lethal to wild type cells. Though their hypermutator phenotype made a detailed global analysis of their genomic changes intractable, important features of the strains' tolerance strategies were revealed via an iModulon analysis of their transcriptome. We discussed mechanisms that involve only 11 mutations and

14 iModulons, but we cover nearly half of all variation in the gene expression of these strains (**Figure 5.2**). The strains use gene expression adaptations to improve tolerance by (i) specializing their stress response by downregulating RpoS and upregulating RpoH, (ii) activating flagellar basal bodies and fimbriae while downregulating FliA with possible effects on the export of heat-damaged, misfolded proteins, (iii) downregulating aerobic metabolism genes to counteract changes to oxygen solubility and autoxidation rates, (iv) upregulating and modifying ionic iron uptake while shifting away from the unnecessary expression of siderophores, and (v) upregulating the previously uncharacterized *yjfIJKL* operon. Each of the changes we describe is supported by coherent mutational mechanisms and existing literature. Together, these adaptations represent the transcriptional systems biology of a high temperature growth phenotype in *E. coli*.

The five mechanisms described above suggest two general principles for mesophilic microbes growing at high temperatures. The first is to streamline stress responses and metabolism – the strains downregulate the RpoS general stress response and the autoxidation-inducing ArcA regulon using TCA cycle mutations. The shift from Fur-1 to Fur-2 also helps to streamline metabolism by de-emphasizing siderophore pathways, though this may be a wild type phenomenon and not an evolved feature. Secondly, the strains need to deal with protein aggregation that occurs at high temperatures. To do so, they can upregulate the RpoH sigmulon of proteases and chaperones, and they may also use the flagellar basal bodies and fimbriae machinery to export misfolded proteins.

This study and similar work on ROS tolerance (**Chapter 4**) [240] emphasize the value of iModulons for building a multi-level understanding of cellular stress tolerance phenotypes. In both cases, the stress response becomes specialized for the given strain by modifying activity of RpoS while leaving the specific stress regulon (RpoH or SoxS) to function as it does in wild type. Interestingly, both cases also showed a shift toward anaerobiosis and higher preference for Fur-2 ionic iron transport, but with different predicted underlying mechanisms. The rich information gleaned from these experiments and datasets ought to motivate further applications of iModulons for understanding unique strains, which will build up more examples associated with each iModulon and further enrich the field's working understanding TRNs.

A particularly fruitful use of iModulon analysis in this study lies in the use activity phase planes (**Figures 5.3B, 5.4D-E, 5.5B, 5.5F, 5.5H)**. Each figure showed a trend that was observed across the many samples of PRECISE-1K [140], and modifications to that trend resulting from regulatory changes in the evolved strains. This is an example of learning from scale, as the evolved trends could not have been understood without the context of the rest of the dataset.

The goal of this study was to demonstrate the usefulness of iModulons for making valuable predictions about the global behavior of hypermutator strains with a desired phenotype. However, we acknowledge this limits the scope of our results. The mutational mechanisms are predicted based on literature associations of the genes and regulons, as opposed to being individually validated. We rely on prior work in the literature, which allows us to cover more of the global features of the transcriptome in a single manuscript. This approach bears the risk of presenting incorrect conclusions, and thus we encourage future studies to more thoroughly validate the hypotheses presented here using traditional methods.

In addition to its contribution to the understanding of the TRN in high temperatures, we hope that this study can inform applications. Engineering flagellar basal bodies for heterologous protein export is a promising approach [254], and we have implicated mutations in the ATPase genes *fliIJ* in a mechanism that upregulates the export basal body without the wasteful production of other motility proteins. We are also the first to report that high temperatures change the activity of ArcA and Fnr and predict that it is due to their sensitivity to temperature-dependent changes in oxygen – this could also be useful for designing cell factories, in which changes to oxygen and temperature commonly occur, and regulatory effects need to be precisely understood. Also, temperature-tolerant pathogenic *E. coli* strains would be more able to survive fever conditions in a host, so the mutations and mechanisms described here could help to explain the evolution of pathogenic strains.

Here, we presented a global characterization of evolved, high temperature-tolerant strains using an emphasis on the transcriptome as opposed to the genome. Our multi-level approach was effective for predicting new mechanisms of heat tolerance and characterizing unknown genes. Given the availability of large amounts of transcriptomic data and tools like iModulon analysis, we believe that TRN evolution will continue to be elucidated in unprecedented ways.

5.4 Materials & methods

5.4.1 Resource availability

RNA-seq data have been deposited to GEO and are publicly available as of the date of publication, under accession number GSE140478. DNAseq data are available from aledb.org under the project "Hot mutL". iModulons and related data are available from iModulonDB.org under the dataset "*E. coli* PRECISE-1K".

All original code and data to generate figures are available at github.com/SBRG/Hot-ALE, which also links to the alignment, ICA, and iModulon analysis workflows [128]. It has been deposited at Zenodo and is publicly available as of the date of publication.

5.4.2 Microbial strains

The starting strain of the original 42c evolution[244] was *E. coli* K-12 MG1655. Mutations for the evolved strains are listed on aledb.org.

5.4.3 Culture conditions

All strains were grown and evolved in M9 minimal medium prepared by addition of 0.1 mM $CaCl_2$, 2 mM $MgSO_4$, 1x trace elements solution, 1x M9 salt solution, and 4 g/L D-glucose to Milli-Q water. The M9 salt solution was composed of 68 g/L Na_2HPO_4, 30 g/L KH_2PO_4, 5 g/L NaCl, and 10 g/L

NH_4Cl. The trace elements solution was prepared by mixing 27 g/L $FeCl_3 \cdot 6 H_2O$, 1.3 g/L $ZnCl_2$, 2 g/L $CoCl_2 \cdot 6 H_2O$, 2 g/L $Na_2MoO_4 \cdot 2 H_2O$, 0.75 g/L $CaCl_2$, 0.91 g/L $CuCl_2$, and 0.5 g/L H_3BO_3 in a Milli-Q water solution consisting of 10% concentrated HCl by final volume. Sterilization was achieved in all solutions and media by filtration through a 0.22 μM PVDF membrane.

5.4.4 Adaptive laboratory evolution

Stage I of the ALE experiment was started from isolates of the wild-type *E. coli* K-12 MG1655, and evolved at 42°C as described previously [244]. Clones were isolated from ten populations at the end of this experiment, and eight of them with distinct mutational histories were used to start the Stage II ALE experiment. Unfortunately, a contamination event early in the Stage II ALE led to the 42c_3 strain becoming the dominant strain in all flasks that were subsequently analyzed.

All cultures during the Stage II evolution were grown in 35 mL flasks with a 15 mL working volume, and were vigorously stirred at 1100 rpm to create a well-mixed and aerobic environment. Initial temperatures for these cultures were set to 42°C. The temperatures were increased by 0.5°C approximately every 150 generations (~15 passages) to give the cultures time to optimize their growth under the new conditions. Due to the higher stress levels, temperature increases were only 0.25°C above 44°C. An automated system was used to propagate the evolving populations over the course of the ALE. To maintain the evolving population at the exponential growth phase, their growth was periodically monitored by taking optical density measurements at a 600 nm wavelength (OD600) on a Tecan Sunrise reader plate (**Figure C.1**). Once reaching the target OD600~0.3 (~1 on a 1 cm path length spectrophotometer), approximately 0.66% of the cells in a population were passaged to the fresh medium. Population samples along the adaptive trajectories were taken by mixing 800 μL of culture with 800 μL of 50% glycerol, and stored at -80°C for subsequent analysis (not reported).

5.4.5 DNA sequencing and mutation calling

Growth-improved clones along the ALE trajectory were isolated and grown in the standard medium condition. Cells were then harvested while in exponential growth and genomic DNA was extracted using a KingFisher Flex Purification system previously validated for the high throughput platform mentioned below [268]. Shotgun metagenomic sequencing libraries were prepared using a miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems). DNA extracts were normalized to 5 ng total input per sample using an Echo 550 acoustic liquid handling robot (Labcyte Inc), and 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions carried out using a Mosquito HTS liquid-handling robot (TTP Labtech Inc). Sequencing adapters were based on the iTru protocol [269], in which short universal adapter stubs are ligated first and then sample-specific barcoded sequences added in a subsequent PCR step. Amplified and barcoded libraries were then quantified using a PicoGreen assay and pooled in approximately equimolar ratios before being sequenced on an Illumina HiSeq 4000 instrument.

Sequencing reads were filtered and trimmed using AfterQC version 0.9.7 [231]. We mapped reads to the *E. coli* K-12 MG1655 reference genome (NC_00913.3) using the breseq pipeline version 0.33.1 [232]. Mutation analysis was performed using ALEdb [146].

5.4.6 Physiological characterization

Cultures were initially inoculated from -80°C glycerol stocks, and grown at 37°C overnight. Physiological adaptation was achieved by growing cell cultures exponentially over 2 passages for 5 to 10 generations at the target temperature for phenotypic characterization. Next, cultures growing at the exponential growth phase were passaged to a 15 mL working volume tube and grown fully aerated. Spectrophotometer readings at OD600 were periodically taken (Thermo Fisher Scientific, Waltham, MA) until stationary phase was reached. Growth rates were determined for each culture by least-squares linear regression of ln(OD600) versus time.

Samples were filtered through a 0.22 micrometer filter (MilliporeSigma, Burlington, MA) at the same time OD600 measurements were taken, and the filtrate was analyzed for glucose and acetate concentrations using a high-performance liquid chromatography system (Agilent Technologies, Santa Clara, CA) with an Aminex HPX-87H column (Bio-Rad Laboratories, Hercules, CA). Glucose uptake rates and acetate production rates in exponential growth were determined by best-fit linear regression of glucose and acetate concentrations versus cell dry weights, multiplied by growth rates over the same sample range. The above described phenotypic characterizations were performed for two biological replicates of each of the selected clonal isolates along the ALE trajectory, at 30°C, 37°C, and 44°C, respectively.

5.4.7 RNA sequencing

During phenotypic characterization, 3 mL of cell broth was taken at OD600~0.6, and immediately added to Qiagen RNAprotect Bacteria Reagent (6 mL). Then, the sample was vortexed for 5 seconds, incubated at room temperature for 5 minutes, and immediately centrifuged for 10 minutes at 5000g. The supernatant was decanted, and the cell pellet was stored in the -80°C. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37°C. Total RNA was isolated and purified using the RNeasy Plus Mini Kit (Qiagen) columns following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. The rRNA was removed using Illumina Ribo-Zero rRNA Removal Kit (Gram-Negative Bacteria). The quantity was determined by Nanodrop 1000 spectrophotometer (Thermo Scientific). The quality was checked using RNA 6000 Pico Kit using Agilent 2100 Bioanalyzer (Agilent). Paired-end, strand-specific RNA-seq library was built with the KAPA RNA Hyper Prep kit (Kapa Biosystems) following manufacturer’s instructions. Libraries were sequenced on an Illumina HiSeq 4000 instrument.

As part of the PRECISE-1K dataset [140], transcriptomic reads were mapped using our pipeline (https://github.com/avsastry/modulome-workflow) [128] and run on Amazon Web Services Batch. First, raw read trimming was performed using Trim Galore with default options, followed by FastQC on the trimmed reads. Next, reads were aligned to the *E. coli* K-12 MG1655 reference genome (NC_000913.3) using Bowtie [233]. The read direction was inferred using RSeQC[234]. Read counts were generated using featureCounts [235]. All quality control metrics were compiled using MultiQC[236]. Finally, the expression dataset was reported in units of log-transformed transcripts per million (log(TPM)).

All included samples passed rigorous quality control, with "high-quality" defined as (i) passing the following FastQC checks: *per_base_sequence_quality, per_sequence_quality_scores, per_base_n_content, adaptor content;* (ii) having at least 500,000 reads mapped to the coding sequences of the reference genome (NC_000913.3); (iii) not being an outlier in a hierarchical clustering based on pairwise Pearson correlation between all samples in PRECISE-1K; and (iv) having a minimum Pearson correlation between biological replicates of 0.95.

5.4.8 iModulon computation and curation

The full PRECISE-1K compendium, including the samples for this study, was used to compute iModulons using our previously described method [32], [140]. The log(TPM) dataset **X** was first centered such that wild-type *E. coli* MG1655 samples in M9 minimal media with glucose had mean expression values of 0 for all genes. Independent component analysis was performed using the Scikit-Learn (v0.19.0) implementation of FastICA [24]. We performed 100 iterations of the algorithm across a range of dimensionalities, and for each dimensionality we pooled and clustered the components with DBSCAN to find robust components which appeared in more than 50 of the iterations. If the dimensionality parameter is too high, ICA will begin to return single gene components; if it is too low, the components will be too dense to represent biological signals. Therefore, we selected a dimensionality which was as high as possible without creating many single gene components, as described [32]. At the optimal dimensionality, the total number of iModulons was 201. The output is composed of matrices **M** [genes x iModulons], which defines the relationship between each iModulon and each gene, and **A** [iModulons x samples], which contains the activity levels for each iModulon in each sample.

For each iModulon, a threshold must be drawn in the **M** matrix to determine which genes are members of each iModulon. These thresholds are based on the distribution of gene weights. The highest weighted genes were progressively removed until the remaining weights had a D'agostino K^2 normality below 550. Thus, the iModulon member genes are outliers from an otherwise normal distribution. iModulon annotation and curation was performed by comparing them against the known TRN from RegulonDB [11]. Names, descriptions, and statistics for each iModulon are available from the PRECISE-1K manuscript [140] and iModulonDB [21]**.**

5.4.9 Differential iModulon activity analysis

DiMAs were calculated as previously described [19], [128]. For each iModulon, a null distribution was generated by calculating the absolute difference between each pair of biological replicates and fitting a log-normal distribution to them. For the groups being compared, their mean difference for each iModulon was compared to that iModulon's null distribution to obtain a p-value. The set of p-values for all iModulons was then false discovery rate (FDR) corrected to generate q-values. Activities were considered significant if they passed an absolute difference threshold of 5 and an FDR of 0.1. The main comparison in this study was between the wild type strain at 42°C (n = 1) and the combined set of all fully evolved strains at 44°C (n = 12). This comparison is shown in **Figure 5.1F**, and its p-values are reported in figure captions throughout the chapter. We used the same statistical algorithm to compare the evolved strains at 30°C (n = 12) and 44°C (n = 12) in **Figure 5.1G**.

5.4.10 iModulon explained variance calculation

The explained variance for each iModulon in this study was calculated using our workflow [128]. Since iModulons are built on a matrix decomposition, the contribution of each one to the overall expression dataset can be calculated. For each iModulon, the column of **M** and the row of **A** for the evolved samples in this study were multiplied together, and the explained variance between the result and the full expression dataset was computed. These explained variance scores were used to size the subsets of the treemap in **Figure 5.1D** and the icons in the third column of **Figure 5.2**. Note that the variance explained by ICA is 'knowledge-based' in contrast to the 'statistic-based' variance explanation provided by the commonly used principal component analysis (PCA).

5.5 Acknowledgements

We thank Amitesh Anand, Xin Fang, and Patrick V. Phaneuf for helpful discussions. This work was funded by National Institutes of Health grants (awards GM102098 and GM057089), and the Novo Nordisk Foundation (awards NNF10CC1016517 and NNF20CC0035580). This research used resources of the National Energy Research Scientific Computing Center, supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions: K.R., K.C., A.M.F., and B.O.P. designed the study. K.C., C.A.O., T.E.S., Y.G., S.X., Y.H., and R.S. performed experiments. A.P. performed simulations. K.R. and K.C. analyzed the data and wrote the manuscript, with contributions from all co-authors.

Chapter 5, in part, is currently being prepared for submission to publication: Rychel K, Chen K, Patel A, Olson CA, Sandberg TE, Gao Y, Xu S, Hefner Y, Szubin R, Feist AM, Palsson BO. Laboratory evolution reveals transcriptional mechanisms underlying thermal adaptation of *Escherichia coli*. 2023. The dissertation author is the primary author.

Chapter 6. Conclusions

6.1 Dissertation Summary

In this dissertation, we have established iModulons as an excellent tool for transcriptomic analysis. In **Chapter 1**, we described how the emergence of genome-scale technologies and rapid accumulation of large biological datasets has heralded a new age of biology. We also identified important gaps in the existing analysis methods for transcriptomic data, which were holding back the field's ability to interpret the sensory and regulatory systems within cells. An important proof-of-concept study by Sastry et al [19] developed iModulon analysis and demonstrated its potential for solving several of the problems in transcriptomics, but more efforts were needed to further develop this tool, broadly apply it, and enable the research community to use it.

In **Chapter 2**, we applied iModulon analysis to characterize the TRN of *Bacillus subtilis*. We globally summarized the major signals underlying diverse conditions, obtained specific hypotheses (some of which were later validated by a collaborator [108]), and described broad re-allocation in sporulation in three phases with a small number of iModulons. Together with the original *E. coli* paper, this work suggested broad applicability of iModulon analysis across the phylogenetic tree, motivating the following chapter.

In **Chapter 3**, we developed and expanded iModulonDB. iModulonDB is a web tool for browsing, searching, visualizing and downloading iModulon-related data and curated knowledge. As of this writing, it has been filled with 16 datasets across 11 diverse species, and it receives over 300 unique users each month. Such rapid growth and adoption speaks to the value of iModulon analysis, and will surely enable the elucidation of further new insights.

In **Chapters 4** and **5**, we moved beyond simply characterizing iModulons by applying and integrating them with other approaches to obtain systems-level understanding of new strains. iModulons proved to be highly informative for this purpose. We characterized laboratory evolved strains which tolerated very high levels of the oxidative stressor paraquat (**Chapter 4**) and thermal stress from high temperatures (**Chapter 5**). For both chapters, we integrated results from genomics and physiological experiments to show relationships between mutations, energy balance, and the transcriptome in detailed knowledge graphs. For **Chapter 4**, we also used a metabolic model to interpret the adaptations in the strains. These chapters listed key stress tolerance strategies which could be used in biomanufacturing for improving stress tolerance of production strains. They also provide examples of multi-omic integration with iModulons, which is important as the field works toward an integrated understanding of genomic, transcriptomic, and metabolomic data.

6.2 Emerging Themes

One of the strengths of iModulon activity analysis, which has emerged as a theme in this dissertation, is that it is bidirectional: iModulon activities have downstream effects on cellular responses, but they also enable inferences about upstream regulatory events. A naive approach to transcriptomic analysis would be to only consider the downstream, e.g. “fumarate reductase is activated, therefore fumarate fermentation is increased”. This reasoning is often flawed, since expression does not necessarily induce function (e.g. there may be no fumarate available to ferment). While it is important to consider that the cell’s phenotypic capacity changes as a result of transcriptional reallocation, it has often been more fruitful to make inferences *upstream* of gene expression. iModulon analysis facilitates upstream interpretation by nature of grouping coregulated genes: the example statement above becomes “the Fnr iModulons are activated, therefore Fnr should be more active”. From there, we can predict reasons that the Fnr TF’s activity increased, for instance because the redox ratio has shifted (see **Section 4.2.11.1** and **5.2.6**). In this way, we can interpret sensory systems and characterize the cell more deeply. When faced with a DiMA, a researcher should consider both the upstream regulation and the downstream effects. Feedback loops wherein an iModulon regulates itself are also common, so it is important to read literature surrounding any regulator carefully.

Compared to typical journal articles in microbiology, the chapters of this dissertation cover a wider breadth of topics at the expense of detailed validation experiments in the wet lab. This is by design: the goal was to gain a global understanding of cellular states by probing each of the major signals in the transcriptome. Where possible, we connect our findings to the broader literature, and prefer to support our predictions with past published work rather than repeat past experiments in our new strains. We are careful to point out what is expected, unexpected, or unexplored, and we encourage detailed follow-up experiments. Results of such experiments will be more easily interpretable in light of the global overviews presented here. We do acknowledge that this type of study design may lead to the proposal of some incorrect hypotheses. Nonetheless, we believe that leveraging and coherently inter-relating decades of past research to achieve a global perspective has value and novelty.

We strongly value FAIR principles – that is, making data and knowledge *f*indable, *a*ccessible, *i*nteroperable, and *r*eusable [270]. The tsunami of FAIR transcriptomic data and the ability to reuse it were the critical advances which enabled this work. The creation of iModulonDB in **Chapter 2** returns this favor to the research community: the data can now be reused again, but with added confidence from our quality checks, with added context from the scale of our analyses, and enriched with added knowledge from the iModulon curation and interpretation process. iModulonDB makes iModulons findable, accessible, and reusable. Interoperability remains a major goal of future efforts, but the relationships between data types demonstrated in **Chapters 4** and **5** are a step in this direction.

6.3 Unreasonable Effectiveness of iModulons & Outlook

This dissertation clearly demonstrated that iModulon analysis has wide applicability as a tool, but we did not further explore the underlying theory. Why is this particular machine learning method so good at unraveling the secrets within transcriptomic datasets?

A similar, but broader question was asked in 1960 by Nobel Prize winning physicist Eugene

Wigner in his work, “The Unreasonable Effectiveness of Mathematics in the Natural Sciences” [271]. In it, he discusses how mathematics seemed at first to simply be a tool and language with which to describe physics. In mathematics, structures like imaginary numbers were used to demonstrate a mathematician’s ingenuity and reasoning, but were never expected to be fundamental to aspects of the real world. However, as physics progressed and took advantage of more advanced mathematical concepts, they proved to be necessary for the formulation of the laws of quantum mechanics. It seems as though math is not simply a tool, but rather an underlying truth which is actually more fundamental than physics.

This analogy is not perfect, since we know that iModulons are not the actual underlying truth of the TRN. They have several key limitations, including their linearity, independence assumption, and sensitivity to arbitrary parameters (see **Section 1.3.3**). Despite this, they are indeed unreasonably effective, as shown by the cases described in this dissertation. Therefore, there ought to be an underlying truth which is shared by real transcriptomes and iModulon analysis.

We speculate that perhaps the key to understanding the transcriptome lies in the separation of structure from activity. With the core matrix decomposition equation **X = M * A,** combined with the flexible statistical constraints for selecting components and our modifications to enhance robustness and tune parameters, ICA is able to obtain a good approximation of both. The genome encodes **M**, which is some relation between regulators and genes arising from promoter sequences, TF binding, RNA degradation, and other interactions. The cellular state and environment provides **A**, some set of quantifiable sensory inputs which exert influence over the regulators. After years of work with iModulons, we believe that other frameworks, such as clustering, which are not formulated around these hidden regulatory variables, are not likely to match prokaryotic biological reality as well as iModulon analysis. However, **M** and **A** could take on forms other than linear matrices.

Perhaps other yet-to-be-developed factorization methods could improve upon this method by finding ways of eliminating some of its assumptions. In addition, the integration of more data types (such as sequence, binding, or modification information) could improve the formulation of **M**. Alternatively, **M** could be built from the bottom-up using detailed promoter models, which would allow interactions to be more explicitly encoded. There is a tradeoff to incorporating this information: understanding of model organisms may improve, but the wide applicability to less characterized organisms (e.g. **Section 3.6**) would be lost.

Even if **M** and **A** are reformulated with more sophisticated algorithms, this iModulon work would still be useful for several reasons. Firstly, it shows the value and applicability of the **M** and **A** framework, which made systems-level understanding of transcriptomes possible. Second, the datasets in iModulonDB provide an excellent resource for testing new methods of TRN inference, since to our knowledge there is no other database of quality controlled, consistently processed, ready-to-use, and knowledge enriched transcriptomic data. Finally, the insights obtained herein are backed by literature, comparisons with other data types, and follow-up experiments, so future decompositions should recapitulate the same insights – this would be another way of testing new **M** and **A** formulations.

6.4 Closing Remarks

In our quest to understand living systems, we are faced with the challenge of elucidating the TRN which senses and adapts to environments at the cellular level. Though the complete output of the TRN could be measured more than two decades ago, interpretability has lagged because of its complexity. In recent years, an unprecedented opportunity has arisen: we can now analyze large compendia of transcriptomic data with machine learning. iModulon analysis has emerged as an effective tool for unraveling these datasets and interpreting the secrets hidden within them. In this dissertation, we developed iModulon analysis by: characterizing an important model organism, disseminating iModulons on an expansive online knowledgebase, and revealing mechanisms of their evolution. The specific insights we gain elucidate new information about a wide range of cellular systems, which may be applicable to the development of biomanufacturing strains and the study of pathogens. This work is a blueprint for interpreting the TRN in an exciting new light.

References

[1] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, no. 5223, pp. 496–512, Jul. 1995, doi: 10.1126/science.7542800.

[2] J. S. Edwards and B. O. Palsson, "Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype *," *J. Biol. Chem.*, vol. 274, no. 25, pp. 17410–17416, Jun. 1999, doi: 10.1074/jbc.274.25.17410.

[3] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nat. Genet.*, vol. 21, no. 1, Art. no. 1, Jan. 1999, doi: 10.1038/4462.

[4] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nat. Biotechnol.*, vol. 28, no. 3, pp. 245–248, Mar. 2010, doi: 10.1038/nbt.1614.

[5] Y. Seif and B. Ø. Palsson, "Path to improving the life cycle and quality of genome-scale models of metabolism," *Cell Syst.*, vol. 12, no. 9, pp. 842–859, Sep. 2021, doi: 10.1016/j.cels.2021.06.005.

[6] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Current status and applications of genome-scale metabolic models," *Genome Biol.*, vol. 20, no. 1, p. 121, Jun. 2019, doi: 10.1186/s13059-019-1730-3.

[7] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[8] D. F. Browning and S. J. W. Busby, "Local and global regulation of transcription initiation in bacteria," *Nat. Rev. Microbiol.*, vol. 14, no. 10, Art. no. 10, Oct. 2016, doi: 10.1038/nrmicro.2016.103.

[9] K. S. Myers, D. M. Park, N. A. Beauchene, and P. J. Kiley, "Defining bacterial regulons using ChIP-seq," *Methods*, vol. 86, pp. 80–88, Sep. 2015, doi: 10.1016/j.ymeth.2015.05.022.

[10] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, A. Kothari, I. M. Keseler, M. Krummenacker, P. E. Midford, Q. Ong, W. K. Ong, S. M. Paley, and P. Subhraveti, "The BioCyc collection of microbial genomes and metabolic pathways," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1085–1093, Jul. 2019, doi: 10.1093/bib/bbx085.

[11] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeida, J. S. García-Sotelo, K. Alquicira-Hernández, L. J. Muñiz-Rascado, P. Peña-Loredo, C. Ishida-Gutiérrez, D. A. Velázquez-Ramírez, V. Del Moral-Chávez, C. Bonavides-Martínez, C.-F. Méndez-Cruz, J. Galagan, and J. Collado-Vides, "RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D212–D220, 08 2019, doi: 10.1093/nar/gky1077.

[12] B. Zhu and J. Stülke, "SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D743–D748, 04 2018, doi: 10.1093/nar/gkx908.

[13] J. S. Shaik and M. Yeasin, “A unified framework for finding differentially expressed genes from microarray experiments,” *BMC Bioinformatics*, vol. 8, no. 1, p. 347, Sep. 2007, doi: 10.1186/1471-2105-8-347.

[14] S. J. Larsen, R. Röttger, H. H. H. W. Schmidt, and J. Baumbach, “E. coli gene regulatory networks are inconsistent with gene expression data,” *Nucleic Acids Res.*, vol. 47, no. 1, pp. 85–92, Jan. 2019, doi: 10.1093/nar/gky1176.

[15] X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J. T. Yurkovich, C. J. Lloyd, Y. Gao, L. Yang, and B. O. Palsson, “Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities,” *Proc. Natl. Acad. Sci.*, Aug. 2017, doi: 10.1073/pnas.1702581114.

[16] R. De Smet and K. Marchal, “Advantages and limitations of current network inference methods,” *Nat. Rev. Microbiol.*, vol. 8, no. 10, Art. no. 10, Oct. 2010, doi: 10.1038/nrmicro2419.

[17] W. Saelens, R. Cannoodt, and Y. Saeys, “A comprehensive evaluation of module detection methods for gene expression data,” *Nat. Commun.*, vol. 9, 2018, doi: 10.1038/s41467-018-03424-4.

[18] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, and C. O’Sullivan, “The Sequence Read Archive: a decade more of explosive growth,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D387–D390, Jan. 2022, doi: 10.1093/nar/gkab1053.

[19] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson, “The Escherichia coli transcriptome mostly consists of independently regulated modules,” *Nat. Commun.*, vol. 10, Dec. 2019, doi: 10.1038/s41467-019-13483-w.

[20] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, Jun. 2000, doi: 10.1016/S0893-6080(00)00026-5.

[21] K. Rychel, K. Decker, A. V. Sastry, P. V. Phaneuf, S. Poudel, and B. O. Palsson, “iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning,” *Nucleic Acids Res.*, 2020, doi: 10.1093/nar/gkaa810.

[22] P. Comon, “Independent component analysis, A new concept?,” *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994, doi: 10.1016/0165-1684(94)90029-9.

[23] J. Shlens, “A Tutorial on Independent Component Analysis,” *ArXiv14042986 Cs Stat*, Apr. 2014, Accessed: Jan. 25, 2019. [Online]. Available: http://arxiv.org/abs/1404.2986

[24] F. Pedregosa, “Scikit-learn: Machine Learning in Python,” *Mach. Learn. PYTHON*, p. 6.

[25] R. A. Saleh and A. K. E. Saleh, “Statistical Properties of the log-cosh Loss Function Used in Machine Learning”.

[26] J. Shlens, “A Tutorial on Principal Component Analysis,” *arXiv.org*, Apr. 03, 2014. https://arxiv.org/abs/1404.1100v1 (accessed Apr. 11, 2023).

[27] K.-K. Yan, G. Fang, N. Bhardwaj, R. P. Alexander, and M. Gerstein, “Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 20, pp. 9186–9191, May 2010, doi: 10.1073/pnas.0914771107.

[28] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol.*, vol. 15, no. 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[29] Z. B. Abrams, T. S. Johnson, K. Huang, P. R. O. Payne, and K. Coombes, “A protocol to evaluate RNA sequencing normalization methods,” *BMC Bioinformatics*, vol. 20, no. 24, p. 679, Dec. 2019, doi: 10.1186/s12859-019-3247-x.

[30] M. L. Arrieta-Ortiz, C. Hafemeister, A. R. Bate, T. Chu, A. Greenfield, B. Shuster, S. N. Barry, M. Gallitto, B. Liu, T. Kacmarczyk, F. Santoriello, J. Chen, C. D. Rodrigues, T. Sato, D. Z. Rudner, A. Driks, R. Bonneau, and P. Eichenberger, “An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network,” *Mol. Syst. Biol.*, vol. 11, no. 11, p. 839, Nov. 2015, doi: 10.15252/msb.20156236.

[31] Y. Yamanaka, T. Shimada, K. Yamamoto, and A. Ishihama, “Transcription factor CecR (YbiH) regulates a set of genes affecting the sensitivity of Escherichia coli against cefoperazone and chloramphenicol,” *Microbiology*, vol. 162, no. 7, pp. 1253–1264, Jul. 2016, doi: 10.1099/mic.0.000292.

[32] J. L. McConn, C. R. Lamoureux, S. Poudel, B. O. Palsson, and A. V. Sastry, “Optimal dimensionality selection for independent component analysis of transcriptomic data,” *BMC Bioinformatics*, vol. 22, no. 1, p. 584, Dec. 2021, doi: 10.1186/s12859-021-04497-7.

[33] X. W. Zhang, Y. L. Yap, D. Wei, F. Chen, and A. Danchin, “Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis,” *Eur. J. Hum. Genet.*, vol. 13, no. 12, pp. 1303–1311, Dec. 2005, doi: 10.1038/sj.ejhg.5201495.

[34] W. Kong, C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang, “A review of independent component analysis application to microarray gene expression data,” *BioTechniques*, vol. 45, no. 5, p. 501, Nov. 2008, doi: 10.2144/000112950.

[35] J. M. Engreitz, B. J. Daigle, Jr, J. J. Marshall, and R. B. Altman, “Independent component analysis: mining microarray data for fundamental human gene expression modules,” *J. Biomed. Inform.*, vol. 43, no. 6, p. 932, Dec. 2010, doi: 10.1016/j.jbi.2010.07.001.

[36] K. J. Karczewski, M. Snyder, R. B. Altman, and N. P. Tatonetti, “Coherent Functional Modules Improve Transcription Factor Target Identification, Cooperativity Prediction, and Disease Association,” *PLOS Genet.*, vol. 10, no. 2, p. e1004122, Feb. 2014, doi: 10.1371/journal.pgen.1004122.

[37] N. Sompairac, P. V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, E. Barillot, F. Radvanyi, A. Gorban, U. Kairov, and A. Zinovyev, “Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets,” *Int. J. Mol. Sci.*, vol. 20, no. 18, p. 4414, Jan. 2019, doi: 10.3390/ijms20184414.

[38] L. Cantini, U. Kairov, A. de Reyniès, E. Barillot, F. Radvanyi, and A. Zinovyev, “Assessing reproducibility of matrix factorization methods in independent transcriptomes,” *Bioinformatics*, vol. 35, no. 21, p. 4307, Nov. 2019, doi: 10.1093/bioinformatics/btz225.

[39] A. Anand, K. Chen, L. Yang, A. V. Sastry, C. A. Olson, S. Poudel, Y. Seif, Y. Hefner, P. V. Phaneuf, S. Xu, R. Szubin, A. M. Feist, and B. O. Palsson, “Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 50, pp. 25287–25292, Dec. 2019, doi: 10.1073/pnas.1909987116.

[40] A. Anand, K. Chen, E. Catoiu, A. V. Sastry, C. A. Olson, T. E. Sandberg, Y. Seif, S. Xu, R. Szubin, L. Yang, A. M. Feist, and B. O. Palsson, “OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States,” *Mol. Biol. Evol.*, vol. 37, no. 3, pp. 660–667, Mar. 2020, doi: 10.1093/molbev/msz251.

[41] J. Tan, A. V. Sastry, K. S. Fremming, S. P. Bjørn, A. Hoffmeyer, S. Seo, B. G. Voldborg, and B. O. Palsson, “Independent component analysis of E. coli’s transcriptome reveals the cellular processes that respond to heterologous gene expression,” *Metab. Eng.*, Jul. 2020, doi: 10.1016/j.ymben.2020.07.002.

[42] K. S. Choudhary, J. A. Kleinmanns, K. Decker, A. V. Sastry, Y. Gao, R. Szubin, Y. Seif, and B. O. Palsson, “Elucidation of Regulatory Modes for Five Two-Component Systems in Escherichia coli Reveals Novel Relationships,” *mSystems*, vol. 5, no. 6, pp. e00980-20, Nov. 2020, doi: 10.1128/mSystems.00980-20.

[43] A. V. Sastry, N. Dillon, A. Anand, S. Poudel, Y. Hefner, S. Xu, R. Szubin, A. M. Feist, V. Nizet, and B. Palsson, “Machine Learning of Bacterial Transcriptomes Reveals Responses Underlying Differential Antibiotic Susceptibility,” *mSphere*, vol. 6, no. 4, pp. e00443-21, Aug. 2021, doi: 10.1128/mSphere.00443-21.

[44] I. A. Rodionova, Y. Gao, A. Sastry, Y. Hefner, H. G. Lim, D. A. Rodionov, M. H. Saier, and B. O. Palsson, “Identification of a transcription factor, PunR, that regulates the purine and purine nucleoside transporter punC in E. coli,” *Commun. Biol.*, vol. 4, no. 1, p. 991, Aug. 2021, doi: 10.1038/s42003-021-02516-0.

[45] A. V. Sastry, A. Hu, D. Heckmann, S. Poudel, E. Kavvas, and B. O. Palsson, “Independent component analysis recovers consistent regulatory signals from disparate datasets,” *PLOS Comput. Biol.*, vol. 17, no. 2, p. e1008647, Feb. 2021, doi: 10.1371/journal.pcbi.1008647.

[46] I. S. Tan and K. S. Ramamurthi, “Spore formation in Bacillus subtilis,” *Environ. Microbiol. Rep.*, vol. 6, no. 3, pp. 212–225, Jun. 2014, doi: 10.1111/1758-2229.12130.

[47] L. S. Cairns, L. Hobley, and N. R. Stanley-Wall, “Biofilm formation by Bacillus subtilis: new insights into regulatory strategies and assembly mechanisms,” *Mol. Microbiol.*, vol. 93, no. 4, pp. 587–598, Aug. 2014, doi: 10.1111/mmi.12697.

[48] D. Schultz, P. G. Wolynes, E. Ben Jacob, and J. N. Onuchic, “Deciding fate in adverse times: sporulation and competence in Bacillus subtilis,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 50, pp. 21027–21034, Dec. 2009, doi: 10.1073/pnas.0912185106.

[49] Y. Gu, X. Xu, Y. Wu, T. Niu, Y. Liu, J. Li, G. Du, and L. Liu, “Advances and prospects of Bacillus subtilis cellular factories: From rational design to industrial applications,” *Metab. Eng.*, vol. 50, pp. 109–121, Nov. 2018, doi: 10.1016/j.ymben.2018.05.006.

[50] P. Nicolas, U. Mader, E. Dervyn, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, S. Aymerich, D. Becher, P. Bisicchia, E. Botella, O. Delumeau, G. Doherty, E. L. Denham, M. J. Fogg, V. Fromion, A. Goelzer, A. Hansen, E. Hartig, C. R. Harwood, G. Homuth, H. Jarmer, M. Jules, E. Klipp, L. Le Chat, F. Lecointe, P. Lewis, W. Liebermeister, A. March, R. A. T. Mars, P. Nannapaneni, D. Noone, S. Pohl, B. Rinn, F. Rugheimer, P. K. Sappa, F. Samson, M. Schaffer, B. Schwikowski, L. Steil, J. Stulke, T. Wiegert, K. M. Devine, A. J. Wilkinson, J. Maarten van Dijl, M. Hecker, U. Volker, P. Bessieres, and P. Noirot, “Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis,” *Science*, vol. 335, no. 6072, pp. 1103–1106, Mar. 2012, doi:

10.1126/science.1206848.

[51] A. Fadda, A. C. Fierro, K. Lemmens, P. Monsieurs, K. Engelen, and K. Marchal, "Inferring the transcriptional network of Bacillus subtilis," *Mol. Biosyst.*, vol. 5, no. 12, pp. 1840–1852, Nov. 2009, doi: 10.1039/B907310H.

[52] S. A. Leyn, M. D. Kazanov, N. V. Sernova, E. O. Ermakova, P. S. Novichkov, and D. A. Rodionov, "Genomic Reconstruction of the Transcriptional Regulatory Network in Bacillus subtilis," *J. Bacteriol.*, vol. 195, no. 11, pp. 2463–2473, Jun. 2013, doi: 10.1128/JB.00140-13.

[53] J. A. Freyre-González, L. G. Treviño-Quintanilla, I. A. Valtierra-Gutiérrez, R. M. Gutiérrez-Ríos, and J. A. Alonso-Pavón, "Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and Escherichia coli unveiled by the natural decomposition approach," *J. Biotechnol.*, vol. 161, no. 3, pp. 278–286, Oct. 2012, doi: 10.1016/j.jbiotec.2012.03.028.

[54] J. A. Freyre-González, A. M. Manjarrez-Casas, E. Merino, M. Martinez-Nuñez, E. Perez-Rueda, and R.-M. Gutiérrez-Ríos, "Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis," *BMC Syst. Biol.*, vol. 7, no. 1, p. 127, Nov. 2013, doi: 10.1186/1752-0509-7-127.

[55] H. Matsuoka, K. Hirooka, and Y. Fujita, "Organization and function of the YsiA regulon of Bacillus subtilis involved in fatty acid degradation," *J. Biol. Chem.*, vol. 282, no. 8, pp. 5180–5194, Feb. 2007, doi: 10.1074/jbc.M606831200.

[56] S. Tojo, T. Satomura, H. Matsuoka, K. Hirooka, and Y. Fujita, "Catabolite Repression of the Bacillus subtilis FadR Regulon, Which Is Involved in Fatty Acid Catabolism," *J. Bacteriol.*, vol. 193, no. 10, pp. 2388–2395, May 2011, doi: 10.1128/JB.00016-11.

[57] J. M. Escorcia-Rodríguez, A. Tauch, and J. A. Freyre-González, "Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1228–1237, 2020, doi: 10.1016/j.csbj.2020.05.015.

[58] P. Gollnick, "Regulation of the Bacillus subtilis trp operon by an RNA-binding protein," *Mol. Microbiol.*, vol. 11, no. 6, pp. 991–997, 1994, doi: 10.1111/j.1365-2958.1994.tb00377.x.

[59] N. Homann, J. Tillonen, and M. Salaspuro, "Microbially produced acetaldehyde from ethanol may increase the risk of colon cancer via folate deficiency," *Int. J. Cancer*, vol. 86, no. 2, pp. 169–173, Apr. 2000, doi: 10.1002/(sici)1097-0215(20000415)86:2<169::aid-ijc4>3.0.co;2-3.

[60] A. A. Badawy, "Tryptophan metabolism in alcoholism," *Adv. Exp. Med. Biol.*, vol. 467, pp. 265–274, 1999, doi: 10.1007/978-1-4615-4709-9_33.

[61] G. V. Gleissenthall, S. Geisler, P. Malik, G. Kemmler, H. Benicke, D. Fuchs, and S. Mechtcheriakov, "Tryptophan metabolism in post-withdrawal alcohol-dependent patients," *Alcohol Alcohol. Oxf. Oxfs.*, vol. 49, no. 3, pp. 251–255, Jun. 2014, doi: 10.1093/alcalc/agu011.

[62] O. N. Ilinskaya, V. V. Ulyanova, D. R. Yarullina, and I. G. Gataullin, "Secretome of Intestinal Bacilli: A Natural Guard against Pathologies," *Front. Microbiol.*, vol. 8, Sep. 2017, doi: 10.3389/fmicb.2017.01666.

[63] S. Magnúsdóttir, D. Ravcheev, V. de Crécy-Lagard, and I. Thiele, "Systematic genome

assessment of B-vitamin biosynthesis suggests co-operation among gut microbes," *Front. Genet.*, vol. 6, p. 148, 2015, doi: 10.3389/fgene.2015.00148.

[64] R. A. Bender, "Regulation of the Histidine Utilization (Hut) System in Bacteria," *Microbiol. Mol. Biol. Rev.*, vol. 76, no. 3, pp. 565–584, Sep. 2012, doi: 10.1128/MMBR.00014-12.

[65] M. P. Cabral, N. C. Soares, J. Aranda, J. R. Parreira, C. Rumbo, M. Poza, J. Valle, V. Calamia, I. Lasa, and G. Bou, "Proteomic and functional analyses reveal a unique lifestyle for Acinetobacter baumannii biofilms and a key role for histidine metabolism," *J. Proteome Res.*, vol. 10, no. 8, pp. 3399–3417, Aug. 2011, doi: 10.1021/pr101299j.

[66] W. Ding, Y. Zhou, Q. Qu, W. Cui, B. O. God'spower, Y. Liu, X. Chen, M. Chen, Y. Yang, and Y. Li, "Azithromycin Inhibits Biofilm Formation by Staphylococcus xylosus and Affects Histidine Biosynthesis Pathway," *Front. Pharmacol.*, vol. 9, p. 740, 2018, doi: 10.3389/fphar.2018.00740.

[67] Y.-H. Zhou, C.-G. Xu, Y.-B. Yang, X.-X. Xing, X. Liu, Q.-W. Qu, W.-Y. Ding, G. Bello-Onaghise, and Y.-H. Li, "Histidine Metabolism and IGPD Play a Key Role in Cefquinome Inhibiting Biofilm Formation of Staphylococcus xylosus," *Front. Microbiol.*, vol. 9, p. 665, 2018, doi: 10.3389/fmicb.2018.00665.

[68] O. Zafra, M. Lamprecht-Grandío, C. G. de Figueras, and J. E. González-Pastor, "Extracellular DNA Release by Undomesticated Bacillus subtilis Is Regulated by Early Competence," *PLOS ONE*, vol. 7, no. 11, p. e48716, Nov. 2012, doi: 10.1371/journal.pone.0048716.

[69] M. F. Wojciechowski, K. R. Peterson, and P. E. Love, "Regulation of the SOS response in Bacillus subtilis: evidence for a LexA repressor homolog," *J. Bacteriol.*, vol. 173, no. 20, pp. 6489–6498, Oct. 1991, doi: 10.1128/jb.173.20.6489-6498.1991.

[70] N. Au, E. Kuester-Schoeck, V. Mandava, L. E. Bothwell, S. P. Canny, K. Chachu, S. A. Colavito, S. N. Fuller, E. S. Groban, L. A. Hensley, T. C. O'Brien, A. Shah, J. T. Tierney, L. L. Tomm, T. M. O'Gara, A. I. Goranov, A. D. Grossman, and C. M. Lovett, "Genetic Composition of the Bacillus subtilis SOS System," *J. Bacteriol.*, vol. 187, no. 22, pp. 7655–7666, Nov. 2005, doi: 10.1128/JB.187.22.7655-7666.2005.

[71] K. Gozzi, C. Ching, S. Paruthiyil, Y. Zhao, V. Godoy-Carter, and Y. Chai, "Bacillus subtilis utilizes the DNA damage response to manage multicellular development," *Npj Biofilms Microbiomes*, vol. 3, no. 1, pp. 1–7, Mar. 2017, doi: 10.1038/s41522-017-0016-3.

[72] S. B. Guttenplan and D. B. Kearns, "Regulation of flagellar motility during biofilm formation," *FEMS Microbiol. Rev.*, vol. 37, no. 6, pp. 849–871, Nov. 2013, doi: 10.1111/1574-6976.12018.

[73] O. Irazoki, J. Aranda, T. Zimmermann, S. Campoy, and J. Barbé, "Molecular Interaction and Cellular Location of RecA and CheW Proteins in Salmonella enterica during SOS Response and Their Implication in Swarming," *Front. Microbiol.*, vol. 7, 2016, doi: 10.3389/fmicb.2016.01560.

[74] P. Randazzo, A. Aubert-Frambourg, A. Guillot, and S. Auger, "The MarR-like protein PchR (YvmB) regulates expression of genes involved in pulcherriminic acid biosynthesis and in the initiation of sporulation in Bacillus subtilis," *BMC Microbiol.*, vol. 16, no. 1, p. 190, 20 2016, doi: 10.1186/s12866-016-0807-3.

[75] S. Arnaouteli, D. A. Matoz-Fernandez, M. Porter, M. Kalamara, J. Abbott, C. E. MacPhee, F. A. Davidson, and N. R. Stanley-Wall, "Pulcherrimin formation controls growth arrest of the Bacillus subtilis biofilm," *Proc. Natl. Acad. Sci.*, vol. 116, no. 27, pp. 13553–13562, Jul. 2019, doi:

10.1073/pnas.1903982116.

[76] K. M. Devine, “Activation of the PhoPR-Mediated Response to Phosphate Limitation Is Regulated by Wall Teichoic Acid Metabolism in Bacillus subtilis,” *Front. Microbiol.*, vol. 9, p. 2678, 2018, doi: 10.3389/fmicb.2018.02678.

[77] J. A. Garnett, F. Marincs, S. Baumberg, P. G. Stockley, and S. E. V. Phillips, “Structure and function of the arginine repressor-operator complex from Bacillus subtilis,” *J. Mol. Biol.*, vol. 379, no. 2, pp. 284–298, May 2008, doi: 10.1016/j.jmb.2008.03.007.

[78] J. Brill, T. Hoffmann, M. Bleisteiner, and E. Bremer, “Osmotically controlled synthesis of the compatible solute proline is critical for cellular defense of Bacillus subtilis against high osmolarity,” *J. Bacteriol.*, vol. 193, no. 19, pp. 5335–5346, Oct. 2011, doi: 10.1128/JB.05490-11.

[79] K. Tanaka, K. Kobayashi, and N. Ogasawara, “The Bacillus subtilis YufLM two-component system regulates the expression of the malate transporters MaeN (YufR) and YflS, and is essential for utilization of malate in minimal medium,” *Microbiol. Read. Engl.*, vol. 149, no. Pt 9, pp. 2317–2329, Sep. 2003, doi: 10.1099/mic.0.26257-0.

[80] W. Winkler, A. Nahvi, and R. R. Breaker, “Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression,” *Nature*, vol. 419, no. 6910, pp. 952–956, Oct. 2002, doi: 10.1038/nature01145.

[81] D. A. Rodionov, A. G. Vitreschak, A. A. Mironov, and M. S. Gelfand, “Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms,” *J. Biol. Chem.*, vol. 277, no. 50, pp. 48949–48959, Dec. 2002, doi: 10.1074/jbc.M208965200.

[82] A. R. Bate, R. Bonneau, and P. Eichenberger, “Bacillus subtilis Systems Biology: Applications of -Omics Techniques to the Study of Endospore Formation,” *Microbiol. Spectr.*, vol. 2, no. 2, Apr. 2014, doi: 10.1128/microbiolspec.TBS-0019-2013.

[83] J. R. Russell, M. T. Cabeen, P. A. Wiggins, J. Paulsson, and R. Losick, “Noise in a phosphorelay drives stochastic entry into sporulation in Bacillus subtilis,” *EMBO J.*, vol. 36, no. 19, pp. 2856–2869, 02 2017, doi: 10.15252/embj.201796988.

[84] P. Eichenberger, M. Fujita, S. T. Jensen, E. M. Conlon, D. Z. Rudner, S. T. Wang, C. Ferguson, K. Haga, T. Sato, J. S. Liu, and R. Losick, “The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis,” *PLoS Biol.*, vol. 2, no. 10, p. e328, Oct. 2004, doi: 10.1371/journal.pbio.0020328.

[85] S. T. Wang, B. Setlow, E. M. Conlon, J. L. Lyon, D. Imamura, T. Sato, P. Setlow, R. Losick, and P. Eichenberger, “The forespore line of gene expression in Bacillus subtilis,” *J. Mol. Biol.*, vol. 358, no. 1, pp. 16–37, Apr. 2006, doi: 10.1016/j.jmb.2006.01.059.

[86] R. Wu, M. Gu, R. Wilton, G. Babnigg, Y. Kim, P. R. Pokkuluri, H. Szurmant, A. Joachimiak, and M. Schiffer, “Insight into the sporulation phosphorelay: crystal structure of the sensor domain of Bacillus subtilis histidine kinase, KinD,” *Protein Sci. Publ. Protein Soc.*, vol. 22, no. 5, pp. 564–576, May 2013, doi: 10.1002/pro.2237.

[87] H. Gao, X. Jiang, K. Pogliano, and A. I. Aronson, “The E1β and E2 Subunits of the Bacillus subtilis Pyruvate Dehydrogenase Complex Are Involved in Regulation of Sporulation,” *J. Bacteriol.*, vol. 184, no. 10, pp. 2780–2788, May 2002, doi: 10.1128/JB.184.10.2780-2788.2002.

[88] S. Srinivasan, I. D. Vladescu, S. A. Koehler, X. Wang, M. Mani, and S. M. Rubinstein, “Matrix Production and Sporulation in Bacillus subtilis Biofilms Localize to Propagating Wave Fronts,” *Biophys. J.*, vol. 114, no. 6, pp. 1490–1498, Mar. 2018, doi: 10.1016/j.bpj.2018.02.002.

[89] Z. E. V. Phillips and M. A. Strauch, “Bacillus subtilis sporulation and stationary phase gene expression,” *Cell. Mol. Life Sci. CMLS*, vol. 59, no. 3, pp. 392–402, Mar. 2002, doi: 10.1007/s00018-002-8431-9.

[90] I. Budde, “Adaptation of Bacillus subtilis to growth at low temperature: a combined transcriptomic and proteomic appraisal,” *Microbiology*, vol. 152, no. 3, pp. 831–853, Mar. 2006, doi: 10.1099/mic.0.28530-0.

[91] J. Narula, M. Fujita, and O. A. Igoshin, “Functional requirements of cellular differentiation: lessons from Bacillus subtilis,” *Curr. Opin. Microbiol.*, vol. 34, pp. 38–46, 2016, doi: 10.1016/j.mib.2016.07.011.

[92] R. W. Ye, W. Tao, L. Bedzyk, T. Young, M. Chen, and L. Li, “Global Gene Expression Profiles of Bacillus subtilis Grown under Anaerobic Conditions,” *J. Bacteriol.*, vol. 182, no. 16, pp. 4458–4465, Aug. 2000, doi: 10.1128/JB.182.16.4458-4465.2000.

[93] M. Serrano, G. Real, J. Santos, J. Carneiro, C. P. M. Jr, and A. O. Henriques, “A Negative Feedback Loop That Limits the Ectopic Activation of a Cell Type–Specific Sporulation Sigma Factor of Bacillus subtilis,” *PLOS Genet.*, vol. 7, no. 9, p. e1002220, Sep. 2011, doi: 10.1371/journal.pgen.1002220.

[94] V. K. Chary, P. Xenopoulos, and P. J. Piggot, “Expression of the σF-Directed csfB Locus Prevents Premature Appearance of σG Activity during Sporulation of Bacillus subtilis,” *J. Bacteriol.*, vol. 189, no. 23, pp. 8754–8757, Dec. 2007, doi: 10.1128/JB.01265-07.

[95] E. B. Mearls, J. Jackter, J. M. Colquhoun, V. Farmer, A. J. Matthews, L. S. Murphy, C. Fenton, and A. H. Camp, “Transcription and translation of the sigG gene is tuned for proper execution of the switch from early to late gene expression in the developing Bacillus subtilis spore,” *PLOS Genet.*, vol. 14, no. 4, p. e1007350, Apr. 2018, doi: 10.1371/journal.pgen.1007350.

[96] V. M. Deppe, S. Klatte, J. Bongaerts, K.-H. Maurer, T. O’Connell, and F. Meinhardt, “Genetic Control of Amadori Product Degradation in Bacillus subtilis via Regulation of frlBONMD Expression by FrlR ▿ ,” *Appl. Environ. Microbiol.*, vol. 77, no. 9, pp. 2839–2846, May 2011, doi: 10.1128/AEM.02515-10.

[97] J. E. González-Pastor, E. C. Hobbs, and R. Losick, “Cannibalism by sporulating bacteria,” *Science*, vol. 301, no. 5632, pp. 510–513, Jul. 2003, doi: 10.1126/science.1086462.

[98] M. Fujita and R. Losick, “Evidence that entry into sporulation in Bacillus subtilis is governed by a gradual increase in the level and activity of the master regulator Spo0A,” *Genes Dev.*, vol. 19, no. 18, pp. 2236–2244, Sep. 2005, doi: 10.1101/gad.1335705.

[99] S. Martínez-Lumbreras, C. Alfano, N. J. Evans, K. M. Collins, K. A. Flanagan, R. A. Atkinson, E. M. Krysztofinska, A. Vydyanath, J. Jackter, S. Fixon-Owoo, A. H. Camp, and R. L. Isaacson, “Structural and Functional Insights into Bacillus subtilis Sigma Factor Inhibitor, CsfB,” *Struct. England1993*, vol. 26, no. 4, pp. 640-648.e5, Apr. 2018, doi: 10.1016/j.str.2018.02.007.

[100] B. J. Kolodziej and R. A. Slepecky, “TRACE METAL REQUIREMENTS FOR SPORULATION OF BACILLUS MEGATERIUM1,” *J. Bacteriol.*, vol. 88, no. 4, pp. 821–830, Oct. 1964.

[101] B. Henares, S. Kommineni, O. Chumsakul, N. Ogasawara, S. Ishikawa, and M. M. Nakano, “The ResD Response Regulator, through Functional Interaction with NsrR and Fur, Plays Three Distinct Roles in Bacillus subtilis Transcriptional Control,” *J. Bacteriol.*, vol. 196, no. 2, pp. 493–503, Jan. 2014, doi: 10.1128/JB.01166-13.

[102] E. Härtig and D. Jahn, “Regulation of the anaerobic metabolism in Bacillus subtilis,” *Adv. Microb. Physiol.*, vol. 61, pp. 195–216, 2012, doi: 10.1016/B978-0-12-394423-8.00005-6.

[103] N. O. Ali, J. Bignon, G. Rapoport, and M. Debarbouille, “Regulation of the Acetoin Catabolic Pathway Is Controlled by Sigma L in Bacillus subtilis,” *J. Bacteriol.*, vol. 183, no. 8, pp. 2497–2504, Apr. 2001, doi: 10.1128/JB.183.8.2497-2504.2001.

[104] I. Dogsa, M. Brloznik, D. Stopar, and I. Mandic-Mulec, “Exopolymer Diversity and the Role of Levan in Bacillus subtilis Biofilms,” *PLOS ONE*, vol. 8, no. 4, p. e62044, Apr. 2013, doi: 10.1371/journal.pone.0062044.

[105] O. Stempler, A. K. Baidya, S. Bhattacharya, G. B. M. Mohan, E. Tzipilevich, L. Sinai, G. Mamou, and S. Ben-Yehuda, “Interspecies nutrient extraction and toxin delivery between bacteria,” *Nat. Commun.*, vol. 8, no. 1, pp. 1–9, Aug. 2017, doi: 10.1038/s41467-017-00344-7.

[106] R. B. D’Agostino and A. Belanger, “A Suggestion for Using Powerful and Informative Tests of Normality,” *Am. Stat.*, vol. 44, no. 4, pp. 316–321, 1990, doi: 10.2307/2684359.

[107] T. Oliphant, “Python for Scientific Computing,” *Comput. Sci. Eng.*, vol. 9, pp. 10–20, Jun. 2007, doi: 10.1109/MCSE.2007.58.

[108] J. D. Tibocha-Bonilla, C. Zuñiga, A. Lekbua, C. Lloyd, K. Rychel, K. Short, and K. Zengler, “Predicting stress response and improved protein overproduction in Bacillus subtilis,” *NPJ Syst. Biol. Appl.*, vol. 8, no. 1, p. 50, Dec. 2022, doi: 10.1038/s41540-022-00259-0.

[109] J. S. Gunn, “The Salmonella PmrAB regulon: lipopolysaccharide modifications, antimicrobial peptide resistance and more,” *Trends Microbiol.*, vol. 16, no. 6, pp. 284–290, Jun. 2008, doi: 10.1016/j.tim.2008.03.007.

[110] P. S. Novichkov, A. E. Kazakov, D. A. Ravcheev, S. A. Leyn, G. Y. Kovaleva, R. A. Sutormin, M. D. Kazanov, W. Riehl, A. P. Arkin, I. Dubchak, and D. A. Rodionov, “RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria,” *BMC Genomics*, vol. 14, p. 745, Nov. 2013, doi: 10.1186/1471-2164-14-745.

[111] S. J. Larsen, R. Röttger, H. H. H. W. Schmidt, and J. Baumbach, “E. coli gene regulatory networks are inconsistent with gene expression data,” *Nucleic Acids Res.*, vol. 47, no. 1, pp. 85–92, Jan. 2019, doi: 10.1093/nar/gky1176.

[112] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Jan. 2013, doi: 10.1093/nar/gks1193.

[113] R. Leinonen, H. Sugawara, and M. Shumway, “The Sequence Read Archive,” *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D19–D21, Jan. 2011, doi: 10.1093/nar/gkq1019.

[114] R. Margolis, L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, M. Guyer, and E. D. Green,

"The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 6, pp. 957–958, Nov. 2014, doi: 10.1136/amiajnl-2014-002974.

[115] H. S. Rhee and B. F. Pugh, "ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy," *Curr. Protoc. Mol. Biol.*, vol. Chapter 21, p. Unit 21.24, Oct. 2012, doi: 10.1002/0471142727.mb2124s100.

[116] S. Poudel, H. Tsunemoto, Y. Seif, A. V. Sastry, R. Szubin, S. Xu, H. Machado, C. A. Olson, A. Anand, J. Pogliano, V. Nizet, and B. O. Palsson, "Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response," *Proc. Natl. Acad. Sci.*, vol. 117, no. 29, pp. 17228–17239, Jul. 2020, doi: 10.1073/pnas.2008413117.

[117] K. Rychel, A. V. Sastry, and B. O. Palsson, "Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome," *bioRxiv*, p. 2020.04.26.062638, Apr. 2020, doi: 10.1101/2020.04.26.062638.

[118] I. A. Rodionova, Y. Gao, A. Sastry, R. Yoo, D. A. Rodionov, M. H. Saier, and B. Ø. Palsson, "Synthesis of the novel transporter YdhC, is regulated by the YdhB transcription factor controlling adenosine and adenine uptake," *bioRxiv*, p. 2020.05.03.074617, May 2020, doi: 10.1101/2020.05.03.074617.

[119] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, 1999, doi: 10.1109/72.761722.

[120] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, in KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.

[121] A. V. Sastry, A. Hu, D. Heckmann, S. Poudel, E. Kavvas, and B. O. Palsson, "Matrix factorization recovers consistent regulatory signals from disparate datasets," *bioRxiv*, p. 2020.04.26.061978, Apr. 2020, doi: 10.1101/2020.04.26.061978.

[122] I. M. Keseler, A. Mackie, A. Santos-Zavaleta, R. Billington, C. Bonavides-Martínez, R. Caspi, C. Fulcher, S. Gama-Castro, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, M. Peralta-Gil, P. Subhraveti, D. A. Velázquez-Ramírez, D. Weaver, J. Collado-Vides, I. Paulsen, and P. D. Karp, "The EcoCyc database: reflecting new knowledge about Escherichia coli K-12," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D543–D550, Jan. 2017, doi: 10.1093/nar/gkw1003.

[123] S. Fuchs, H. Mehlan, J. Bernhardt, A. Hennig, S. Michalik, K. Surmann, J. Pané-Farré, A. Giese, S. Weiss, L. Backert, A. Herbig, K. Nieselt, M. Hecker, U. Völker, and U. Mäder, "AureoWiki‑The repository of the Staphylococcus aureus research and annotation community," *Int. J. Med. Microbiol. IJMM*, vol. 308, no. 6, pp. 558–568, Aug. 2018, doi: 10.1016/j.ijmm.2017.11.011.

[124] J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, and P. Bork, "Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper," *Mol. Biol. Evol.*, vol. 34, no. 8, pp. 2115–2122, Aug. 2017, doi: 10.1093/molbev/msx148.

[125] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork, "eggNOG 5.0: a

hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D309–D314, Jan. 2019, doi: 10.1093/nar/gky1085.

[126] L. Yang, N. Mih, A. Anand, J. H. Park, J. Tan, J. T. Yurkovich, J. M. Monk, C. J. Lloyd, T. E. Sandberg, S. W. Seo, D. Kim, A. V. Sastry, P. Phaneuf, Y. Gao, J. T. Broddrick, K. Chen, D. Heckmann, R. Szubin, Y. Hefner, A. M. Feist, and B. O. Palsson, "Cellular responses to reactive oxygen species are predicted from molecular mechanisms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 28, pp. 14368–14373, Jul. 2019, doi: 10.1073/pnas.1905039116.

[127] J. T. Brosnan and M. E. Brosnan, "Branched-chain amino acids: enzyme and substrate regulation," *J. Nutr.*, vol. 136, no. 1 Suppl, pp. 207S–11S, 2006, doi: 10.1093/jn/136.1.207S.

[128] A. V. Sastry, S. Poudel, K. Rychel, R. Yoo, C. R. Lamoureux, S. Chauhan, Z. B. Haiman, T. A. Bulushi, Y. Seif, and B. O. Palsson, "Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks." bioRxiv, p. 2021.07.01.450581, Jul. 02, 2021. doi: 10.1101/2021.07.01.450581.

[129] Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, and R. Knight, "Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea," *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Dec. 2019, doi: 10.1038/s41467-019-13443-4.

[130] N. D. Menon, S. Poudel, A. V. Sastry, K. Rychel, R. Szubin, N. Dillon, H. Tsunemoto, R. Yoo, Y. Hirose, P. N. Rather, B. G. Nair, G. B. Kumar, B. O. Palsson, and V. Nizet, "Independent component analysis reveals 49 independently modulated gene sets within the global transcriptional regulatory architecture of multidrug-resistant Acinetobacter baumannii".

[131] R. Yoo, K. Rychel, S. Poudel, T. Al-Bulushi, Y. Yuan, S. Chauhan, C. Lamoureux, B. O. Palsson, and A. Sastry, "Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection," *mSphere*, vol. 7, no. 2, p. e0003322, Apr. 2022, doi: 10.1128/msphere.00033-22.

[132] A. Rajput, H. Tsunemoto, A. V. Sastry, R. Szubin, K. Rychel, J. Sugie, J. Pogliano, and B. O. Palsson, "Machine learning from Pseudomonas aeruginosa transcriptomes identifies independently modulated sets of genes associated with known transcriptional regulators," *Nucleic Acids Res.*, vol. 50, no. 7, pp. 3658–3672, Apr. 2022, doi: 10.1093/nar/gkac187.

[133] A. Rajput, H. Tsunemoto, A. V. Sastry, R. Szubin, K. Rychel, S. M. Chauhan, J. Pogliano, and B. O. Palsson, "Advanced transcriptomic analysis reveals the role of efflux pumps and media composition in antibiotic responses of Pseudomonas aeruginosa," *Nucleic Acids Res.*, vol. 50, no. 17, pp. 9675–9688, Sep. 2022, doi: 10.1093/nar/gkac743.

[134] H. G. Lim, K. Rychel, A. V. Sastry, G. J. Bentley, J. Mueller, H. S. Schindel, P. E. Larsen, P. D. Laible, A. M. Guss, W. Niu, C. W. Johnson, G. T. Beckham, A. M. Feist, and B. O. Palsson, "Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network," *Metab. Eng.*, Apr. 2022, doi: 10.1016/j.ymben.2022.04.004.

[135] Y. Yuan, Y. Seif, K. Rychel, R. Yoo, S. Chauhan, S. Poudel, T. Al-Bulushi, B. O. Palsson, and A. V. Sastry, "Pan-Genome Analysis of Transcriptional Regulation in Six Salmonella enterica Serovar Typhimurium Strains Reveals Their Different Regulatory Structures," *mSystems*, p.

e0046722, Nov. 2022, doi: 10.1128/msystems.00467-22.

[136] S. M. Chauhan, S. Poudel, K. Rychel, C. Lamoureux, R. Yoo, T. Al Bulushi, Y. Yuan, B. O. Palsson, and A. V. Sastry, "Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile Sulfolobus acidocaldarius," *Front. Microbiol.*, vol. 12, 2021, Accessed: Feb. 23, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2021.753521

[137] Y. Hirose, S. Poudel, A. V. Sastry, K. Rychel, R. Szubin, D. Zielinski, H. G. Lim, N. Menon, H. Bergsten, S. Uchiyama, T. Hanada, S. Kawabata, B. O. Palsson, and V. Nizet, "Elucidation of independently modulated genes in Streptococcus pyogenes reveals carbon sources that control its expression of hemolytic toxins." bioRxiv, p. 2022.08.04.502797, Aug. 04, 2022. doi: 10.1101/2022.08.04.502797.

[138] J. Shin, K. Rychel, and B. Palsson, "Systems Biology of Competency in Bacteria is Revealed by Applying Novel Data Analytics to the Transcriptome." Rochester, NY, Dec. 23, 2022. doi: 10.2139/ssrn.4309024.

[139] J. C. Hyun and B. O. Palsson, "Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome".

[140] C. R. Lamoureux, K. T. Decker, A. V. Sastry, K. Rychel, Y. Gao, J. L. McConn, D. C. Zielinski, and B. O. Palsson, "A multi-scale transcriptional regulatory network knowledge base for *Escherichia coli*," Bioinformatics, preprint, Apr. 2021. doi: 10.1101/2021.04.08.439047.

[141] P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, and M. Gerstein, "The real cost of sequencing: scaling computation to keep pace with data generation," *Genome Biol.*, vol. 17, p. 53, Mar. 2016, doi: 10.1186/s13059-016-0917-0.

[142] J. H. Yang, S. N. Wright, M. Hamblin, D. McCloskey, M. A. Alcantar, L. Schrübbers, A. J. Lopatkin, S. Satish, A. Nili, B. O. Palsson, G. C. Walker, and J. J. Collins, "A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action," *Cell*, vol. 177, no. 6, pp. 1649-1661.e9, May 2019, doi: 10.1016/j.cell.2019.04.016.

[143] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics Data Integration, Interpretation, and Its Application," *Bioinforma. Biol. Insights*, vol. 14, p. 1177932219899051, Jan. 2020, doi: 10.1177/1177932219899051.

[144] T. E. Sandberg, M. J. Salazar, L. L. Weng, B. O. Palsson, and A. M. Feist, "The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology," *Metab. Eng.*, vol. 56, pp. 1–16, Dec. 2019, doi: 10.1016/j.ymben.2019.08.004.

[145] G. L. Peabody, J. Winkler, and K. C. Kao, "Tools for developing tolerance to toxic chemicals in microbial systems and perspectives on moving the field forward and into the industrial setting," *Curr. Opin. Chem. Eng.*, vol. 6, pp. 9–17, Nov. 2014, doi: 10.1016/j.coche.2014.08.001.

[146] P. V. Phaneuf, D. Gosting, B. O. Palsson, and A. M. Feist, "ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1164–D1171, Jan. 2019, doi: 10.1093/nar/gky983.

[147] D. Hughes and D. I. Andersson, "Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms," *Nat. Rev. Genet.*, vol. 16, no. 8, pp. 459–471, Aug. 2015,

doi: 10.1038/nrg3922.

[148] K. Rychel, A. V. Sastry, and B. O. Palsson, “Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome,” *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Dec. 2020, doi: 10.1038/s41467-020-20153-9.

[149] A. Anand, C. A. Olson, A. V. Sastry, A. Patel, R. Szubin, L. Yang, A. M. Feist, and B. O. Palsson, “Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications,” *Cell Rep.*, vol. 35, no. 1, p. 108961, Apr. 2021, doi: 10.1016/j.celrep.2021.108961.

[150] A. Anand, A. Patel, K. Chen, C. A. Olson, P. V. Phaneuf, C. Lamoureux, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson, “Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity,” *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41467-022-30877-5.

[151] E. S. Kavvas, C. P. Long, A. Sastry, S. Poudel, M. R. Antoniewicz, Y. Ding, E. T. Mohamed, R. Szubin, J. M. Monk, A. M. Feist, and B. O. Palsson, “Experimental Evolution Reveals Unifying Systems-Level Adaptations but Diversity in Driving Genotypes,” *mSystems*, p. e0016522, Oct. 2022, doi: 10.1128/msystems.00165-22.

[152] J. M. Monk, C. J. Lloyd, E. Brunk, N. Mih, A. Sastry, Z. King, R. Takeuchi, W. Nomura, Z. Zhang, H. Mori, A. M. Feist, and B. O. Palsson, “iML1515, a knowledgebase that computes Escherichia coli traits,” *Nat. Biotechnol.*, vol. 35, no. 10, pp. 904–908, 11 2017, doi: 10.1038/nbt.3956.

[153] K. Chen, Y. Gao, N. Mih, E. J. O’Brien, L. Yang, and B. O. Palsson, “Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation,” *Proc. Natl. Acad. Sci.*, vol. 114, no. 43, pp. 11548–11553, Oct. 2017, doi: 10.1073/pnas.1705524114.

[154] B. Du, L. Yang, C. J. Lloyd, X. Fang, and B. O. Palsson, “Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in Escherichia coli,” *PLOS Comput. Biol.*, vol. 15, no. 12, p. e1007525, Dec. 2019, doi: 10.1371/journal.pcbi.1007525.

[155] H. M. Hassan and I. Fridovich, “Paraquat and Escherichia coli. Mechanism of production of extracellular superoxide radical.,” *J. Biol. Chem.*, vol. 254, no. 21, pp. 10846–10852, Nov. 1979, doi: 10.1016/S0021-9258(19)86598-5.

[156] H. M. Hassan and I. Fridovich, “Intracellular production of superoxide radical and of hydrogen peroxide by redox active compounds,” *Arch. Biochem. Biophys.*, vol. 196, no. 2, pp. 385–395, Sep. 1979, doi: 10.1016/0003-9861(79)90289-3.

[157] J. A. Imlay, “Where in the world do bacteria experience oxidative stress?,” *Environ. Microbiol.*, vol. 21, no. 2, pp. 521–530, Feb. 2019, doi: 10.1111/1462-2920.14445.

[158] M. Fasnacht and N. Polacek, “Oxidative Stress in Bacteria and the Central Dogma of Molecular Biology,” *Front. Mol. Biosci.*, vol. 8, 2021, Accessed: Jul. 05, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmolb.2021.671037

[159] J. A. Imlay, “The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium,” *Nat. Rev. Microbiol.*, vol. 11, no. 7, pp. 443–454, Jul. 2013, doi: 10.1038/nrmicro3032.

[160] J. A. Imlay, “Diagnosing oxidative stress in bacteria: not as easy as you might think,” *Curr. Opin. Microbiol.*, vol. 24, pp. 124–131, Apr. 2015, doi: 10.1016/j.mib.2015.01.004.

[161] M. Schieber and N. S. Chandel, “ROS Function in Redox Signaling and Oxidative Stress,” *Curr. Biol. CB*, vol. 24, no. 10, pp. R453–R462, May 2014, doi: 10.1016/j.cub.2014.03.034.

[162] R. A. LaCroix, T. E. Sandberg, E. J. O’Brien, J. Utrilla, A. Ebrahim, G. I. Guzman, R. Szubin, B. O. Palsson, and A. M. Feist, “Use of Adaptive Laboratory Evolution To Discover Key Mutations Enabling Rapid Growth of Escherichia coli K-12 MG1655 on Glucose Minimal Medium,” *Appl. Environ. Microbiol.*, vol. 81, no. 1, pp. 17–30, Jan. 2015, doi: 10.1128/AEM.02246-14.

[163] R. A. LaCroix, B. O. Palsson, and A. M. Feist, “A Model for Designing Adaptive Laboratory Evolution Experiments,” *Appl. Environ. Microbiol.*, vol. 83, no. 8, pp. e03115-16, Mar. 2017, doi: 10.1128/AEM.03115-16.

[164] L. P. Candeias and S. Steenken, “Electron transfer in di(deoxy)nucleoside phosphates in aqueous solution: rapid migration of oxidative damage (via adenine) to guanine,” *J. Am. Chem. Soc.*, vol. 115, no. 6, pp. 2437–2440, Mar. 1993, doi: 10.1021/ja00059a044.

[165] J. J. Foti, B. Devadoss, J. A. Winkler, J. J. Collins, and G. C. Walker, “Oxidation of the guanine nucleotide pool underlies cell death by bactericidal antibiotics,” *Science*, vol. 336, no. 6079, pp. 315–319, Apr. 2012, doi: 10.1126/science.1219192.

[166] H. Yerushalmi, M. Lebendiker, and S. Schuldiner, “EmrE, an Escherichia coli 12-kDa multidrug transporter, exchanges toxic cations and H+ and is soluble in organic solvents,” *J. Biol. Chem.*, vol. 270, no. 12, pp. 6856–6863, Mar. 1995, doi: 10.1074/jbc.270.12.6856.

[167] Y. Sekine, K. Aihara, and E. Ohtsubo, “Linearization and transposition of circular molecules of insertion sequence IS3,” *J. Mol. Biol.*, vol. 294, no. 1, pp. 21–34, Nov. 1999, doi: 10.1006/jmbi.1999.3181.

[168] N. D. Grindley, “IS1 insertion generates duplication of a nine base pair sequence at its target site,” *Cell*, vol. 13, no. 3, pp. 419–426, Mar. 1978, doi: 10.1016/0092-8674(78)90316-1.

[169] M. M. Klepsch, M. Kovermann, C. Löw, J. Balbach, H. P. Permentier, F. Fusetti, J. W. de Gier, D. J. Slotboom, and R. P.-A. Berntsson, “Escherichia coli peptide binding protein OppA has a preference for positively charged peptides,” *J. Mol. Biol.*, vol. 414, no. 1, pp. 75–85, Nov. 2011, doi: 10.1016/j.jmb.2011.09.043.

[170] P. L. Freddolino, S. Amini, and S. Tavazoie, “Newly identified genetic variations in common Escherichia coli MG1655 stock cultures,” *J. Bacteriol.*, vol. 194, no. 2, pp. 303–306, Jan. 2012, doi: 10.1128/JB.06087-11.

[171] K. Schnetz and B. Rak, “IS5: a mobile enhancer of transcription in Escherichia coli,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 4, pp. 1244–1248, Feb. 1992, doi: 10.1073/pnas.89.4.1244.

[172] S. Mongkolsuk and J. D. Helmann, “Regulation of inducible peroxide stress responses,” *Mol. Microbiol.*, vol. 45, no. 1, pp. 9–15, Jul. 2002, doi: 10.1046/j.1365-2958.2002.03015.x.

[173] S. Barshishat, M. Elgrably-Weiss, J. Edelstein, J. Georg, S. Govindarajan, M. Haviv, P. R. Wright, W. R. Hess, and S. Altuvia, “OxyS small RNA induces cell cycle arrest to allow DNA damage repair,” *EMBO J.*, vol. 37, no. 3, pp. 413–426, Feb. 2018, doi: 10.15252/embj.201797651.

[174] A. Sen and J. A. Imlay, “How Microbes Defend Themselves From Incoming Hydrogen

Peroxide," *Front. Immunol.*, vol. 12, 2021, Accessed: May 18, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2021.667343

[175] A. Bagg and J. B. Neilands, "Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli," *Biochemistry*, vol. 26, no. 17, pp. 5471–5477, Aug. 1987, doi: 10.1021/bi00391a039.

[176] A. T. Smith, R. O. Linkous, N. J. Max, A. E. Sestok, V. A. Szalai, and K. N. Chacón, "The FeoC [4Fe-4S] Cluster Is Redox-Active and Rapidly Oxygen-Sensitive," *Biochemistry*, vol. 58, no. 49, pp. 4935–4949, Dec. 2019, doi: 10.1021/acs.biochem.9b00745.

[177] K. Esquilin-Lebron, S. Dubrac, F. Barras, and J. M. Boyd, "Bacterial Approaches for Assembling Iron-Sulfur Proteins," *mBio*, vol. 12, no. 6, pp. e02425-21, Nov. 2021, doi: 10.1128/mBio.02425-21.

[178] B. Blanc, M. Clémancey, J.-M. Latour, M. Fontecave, and S. Ollagnier de Choudens, "Molecular Investigation of Iron–Sulfur Cluster Assembly Scaffolds under Stress," *Biochemistry*, vol. 53, no. 50, pp. 7867–7869, Dec. 2014, doi: 10.1021/bi5012496.

[179] B. Blanc, C. Gerez, and S. Ollagnier de Choudens, "Assembly of Fe/S proteins in bacterial systems: Biochemistry of the bacterial ISC system," *Biochim. Biophys. Acta*, vol. 1853, no. 6, pp. 1436–1447, Jun. 2015, doi: 10.1016/j.bbamcr.2014.12.009.

[180] P. S. Garcia, S. Gribaldo, B. Py, and F. Barras, "The SUF system: an ABC ATPase-dependent protein complex with a role in Fe-S cluster biogenesis," *Res. Microbiol.*, vol. 170, no. 8, pp. 426–434, Dec. 2019, doi: 10.1016/j.resmic.2019.08.001.

[181] B. Roche, L. Aussel, B. Ezraty, P. Mandin, B. Py, and F. Barras, "Iron/sulfur proteins biogenesis in prokaryotes: Formation, regulation and diversity," *Biochim. Biophys. Acta BBA - Bioenerg.*, vol. 1827, no. 3, pp. 455–469, Mar. 2013, doi: 10.1016/j.bbabio.2012.12.010.

[182] S. Rajagopalan, S. J. Teter, P. H. Zwart, R. G. Brennan, K. J. Phillips, and P. J. Kiley, "Studies of IscR reveal a unique mechanism for metal-dependent regulation of DNA binding specificity," *Nat. Struct. Mol. Biol.*, vol. 20, no. 6, p. 740, Jun. 2013, doi: 10.1038/nsmb.2568.

[183] M. Lénon, R. Arias-Cartín, and F. Barras, "The Fe–S proteome of Escherichia coli: prediction, function, and fate," *Metallomics*, vol. 14, no. 5, p. mfac022, May 2022, doi: 10.1093/mtomcs/mfac022.

[184] D. Touati, "Sensing and protecting against superoxide stress in Escherichia coli--how many ways are there to trigger soxRS response?," *Redox Rep. Commun. Free Radic. Res.*, vol. 5, no. 5, pp. 287–293, 2000, doi: 10.1179/135100000101535825.

[185] H. Rosenberg, R. G. Gerdes, and K. Chegwidden, "Two systems for the uptake of phosphate in Escherichia coli," *J. Bacteriol.*, vol. 131, no. 2, pp. 505–511, Aug. 1977, doi: 10.1128/jb.131.2.505-511.1977.

[186] S. I. Bibikov, R. Biran, K. E. Rudd, and J. S. Parkinson, "A signal transducer for aerotaxis in Escherichia coli," *J. Bacteriol.*, vol. 179, no. 12, pp. 4075–4079, Jun. 1997, doi: 10.1128/jb.179.12.4075-4079.1997.

[187] B. L. Taylor, I. B. Zhulin, and M. S. Johnson, "Aerotaxis and other energy-sensing behavior in bacteria," *Annu. Rev. Microbiol.*, vol. 53, pp. 103–128, 1999, doi:

10.1146/annurev.micro.53.1.103.

[188] B. M. Prüss, J. W. Campbell, T. K. Van Dyk, C. Zhu, Y. Kogan, and P. Matsumura, “FlhD/FlhC is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through induction of the methyl-accepting chemotaxis protein Aer,” *J. Bacteriol.*, vol. 185, no. 2, pp. 534–543, Jan. 2003, doi: 10.1128/JB.185.2.534-543.2003.

[189] C. Pesavento and R. Hengge, “The global repressor FliZ antagonizes gene expression by σS-containing RNA polymerase due to overlapping DNA binding specificity,” *Nucleic Acids Res.*, vol. 40, no. 11, pp. 4783–4793, Jun. 2012, doi: 10.1093/nar/gks055.

[190] S. Gottesman, “Trouble is coming: Signaling pathways that regulate general stress responses in bacteria,” *J. Biol. Chem.*, vol. 294, no. 31, pp. 11685–11700, Aug. 2019, doi: 10.1074/jbc.REV119.005593.

[191] H. E. Schellhorn, “Function, Evolution, and Composition of the RpoS Regulon in Escherichia coli,” *Front. Microbiol.*, vol. 11, 2020, Accessed: Jul. 05, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2020.560099

[192] E. S. Kavvas, M. Antoniewicz, C. Long, Y. Ding, J. M. Monk, B. O. Palsson, and A. M. Feist, “Laboratory evolution of multiple E. coli strains reveals unifying principles of adaptation but diversity in driving genotypes.” bioRxiv, p. 2020.05.19.104992, May 20, 2020. doi: 10.1101/2020.05.19.104992.

[193] J. Utrilla, E. J. O’Brien, K. Chen, D. McCloskey, J. Cheung, H. Wang, D. Armenta-Medina, A. M. Feist, and B. O. Palsson, “Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution,” *Cell Syst.*, vol. 2, no. 4, pp. 260–271, Apr. 2016, doi: 10.1016/j.cels.2016.04.003.

[194] S. E. Irving, N. R. Choudhury, and R. M. Corrigan, “The stringent response and physiological roles of (pp)pGpp in bacteria,” *Nat. Rev. Microbiol.*, pp. 1–16, Nov. 2020, doi: 10.1038/s41579-020-00470-y.

[195] S. Jang and J. A. Imlay, “Micromolar intracellular hydrogen peroxide disrupts metabolism by damaging iron-sulfur enzymes,” *J. Biol. Chem.*, vol. 282, no. 2, pp. 929–937, Jan. 2007, doi: 10.1074/jbc.M607646200.

[196] A. Majumder, M. Fang, K. J. Tsai, C. Ueguchi, T. Mizuno, and H. Y. Wu, “LeuO expression in response to starvation for branched-chain amino acids,” *J. Biol. Chem.*, vol. 276, no. 22, pp. 19046–19051, Jun. 2001, doi: 10.1074/jbc.M100945200.

[197] J. T. Jarrett, “The novel structure and chemistry of iron-sulfur clusters in the adenosylmethionine-dependent radical enzyme biotin synthase,” *Arch. Biochem. Biophys.*, vol. 433, no. 1, pp. 312–321, Jan. 2005, doi: 10.1016/j.abb.2004.10.003.

[198] D. Beckett, “Biotin sensing: universal influence of biotin status on transcription,” *Annu. Rev. Genet.*, vol. 41, pp. 443–464, 2007, doi: 10.1146/annurev.genet.41.042007.170450.

[199] C. A. Mauzy and M. A. Hermodson, “Structural and functional analyses of the repressor, RbsR, of the ribose operon of Escherichia coli,” *Protein Sci. Publ. Protein Soc.*, vol. 1, no. 7, pp. 831–842, Jul. 1992, doi: 10.1002/pro.5560010701.

[200] D. Christodoulou, H. Link, T. Fuhrer, K. Kochanowski, L. Gerosa, and U. Sauer, “Reserve Flux

Capacity in the Pentose Phosphate Pathway Enables Escherichia coli's Rapid Response to Oxidative Stress," *Cell Syst.*, vol. 6, no. 5, pp. 569-578.e7, May 2018, doi: 10.1016/j.cels.2018.04.009.

[201] K. Y. Choi and H. Zalkin, "Structural characterization and corepressor binding of the Escherichia coli purine repressor.," *J. Bacteriol.*, vol. 174, no. 19, pp. 6207–6214, Oct. 1992.

[202] A. Lochowska, R. Iwanicka-Nowicka, D. Plochocka, and M. M. Hryniewicz, "Functional dissection of the LysR-type CysB transcriptional regulator. Regions important for DNA binding, inducer response, oligomerization, and positive control," *J. Biol. Chem.*, vol. 276, no. 3, pp. 2098–2107, Jan. 2001, doi: 10.1074/jbc.M007192200.

[203] J. Ostrowski and N. M. Kredich, "In vitro interactions of CysB protein with the cysJIH promoter of Salmonella typhimurium: inhibitory effects of sulfide," *J. Bacteriol.*, vol. 172, no. 2, pp. 779–785, Feb. 1990, doi: 10.1128/jb.172.2.779-785.1990.

[204] L. I. Leichert, F. Gehrke, H. V. Gudiseva, T. Blackwell, M. Ilbert, A. K. Walker, J. R. Strahler, P. C. Andrews, and U. Jakob, "Quantifying changes in the thiol redox proteome upon oxidative stress in vivo," *Proc. Natl. Acad. Sci.*, vol. 105, no. 24, pp. 8197–8202, Jun. 2008, doi: 10.1073/pnas.0707723105.

[205] I. Ohtsu, N. Wiriyathanawudhiwong, S. Morigasaki, T. Nakatani, H. Kadokura, and H. Takagi, "The l-Cysteine/l-Cystine Shuttle System Provides Reducing Equivalents to the Periplasm in Escherichia coli," *J. Biol. Chem.*, vol. 285, no. 23, pp. 17479–17487, Jun. 2010, doi: 10.1074/jbc.M109.081356.

[206] C. Rensing and G. Grass, "Escherichia coli mechanisms of copper homeostasis in a changing environment," *FEMS Microbiol. Rev.*, vol. 27, no. 2–3, pp. 197–213, Jun. 2003, doi: 10.1016/S0168-6445(03)00049-4.

[207] L. Benov, H. Sage, and I. Fridovich, "The copper- and zinc-containing superoxide dismutase from Escherichia coli: molecular weight and stability," *Arch. Biochem. Biophys.*, vol. 340, no. 2, pp. 305–310, Apr. 1997, doi: 10.1006/abbi.1997.9940.

[208] C. W. Hill, J. A. Gray, and H. Brody, "Use of the isocitrate dehydrogenase structural gene for attachment of e14 in Escherichia coli K-12," *J. Bacteriol.*, vol. 171, no. 7, pp. 4083–4084, Jul. 1989, doi: 10.1128/jb.171.7.4083-4084.1989.

[209] T. Fukushima, K. Yamada, A. Isobe, K. Shiwaku, and Y. Yamane, "Mechanism of cytotoxicity of paraquat. I. NADH oxidation and paraquat radical formation via complex I," *Exp. Toxicol. Pathol. Off. J. Ges. Toxikol. Pathol.*, vol. 45, no. 5–6, pp. 345–349, Oct. 1993, doi: 10.1016/S0940-2993(11)80424-0.

[210] S. I. Liochev and I. Fridovich, "Paraquat diaphorases in Escherichia coli," *Free Radic. Biol. Med.*, vol. 16, no. 5, pp. 555–559, May 1994, doi: 10.1016/0891-5849(94)90055-8.

[211] H. Shimada, K. Hirai, E. Simamura, and J. Pan, "Mitochondrial NADH-quinone oxidoreductase of the outer membrane is responsible for paraquat cytotoxicity in rat livers," *Arch. Biochem. Biophys.*, vol. 351, no. 1, pp. 75–81, Mar. 1998, doi: 10.1006/abbi.1997.0557.

[212] K. Chen, A. Anand, C. Olson, T. E. Sandberg, Y. Gao, N. Mih, and B. O. Palsson, "Bacterial fitness landscapes stratify based on proteome allocation associated with discrete aero-types," *PLOS Comput. Biol.*, vol. 17, no. 1, p. e1008596, Jan. 2021, doi: 10.1371/journal.pcbi.1008596.

[213] G. Korkmaz, M. Holm, T. Wiens, and S. Sanyal, “Comprehensive Analysis of Stop Codon Usage in Bacteria and Its Correlation with Release Factor Abundance,” *J. Biol. Chem.*, vol. 289, no. 44, pp. 30334–30342, Oct. 2014, doi: 10.1074/jbc.M114.606632.

[214] M. A. Rould, J. J. Perona, D. Söll, and T. A. Steitz, “Structure of E. coli Glutaminyl-tRNA Synthetase Complexed with tRNAGln and ATP at 2.8 Å Resolution,” *Science*, vol. 246, no. 4934, pp. 1135–1142, Dec. 1989, doi: 10.1126/science.2479982.

[215] G. Eggertsson and D. Söll, “Transfer ribonucleic acid-mediated suppression of termination codons in Escherichia coli.,” *Microbiol. Rev.*, vol. 52, no. 3, pp. 354–374, Sep. 1988.

[216] S. I. Liochev, A. Hausladen, W. F. Beyer, and I. Fridovich, “NADPH: ferredoxin oxidoreductase acts as a paraquat diaphorase and is a member of the soxRS regulon.,” *Proc. Natl. Acad. Sci.*, vol. 91, no. 4, pp. 1328–1331, Feb. 1994, doi: 10.1073/pnas.91.4.1328.

[217] U. Sauer, F. Canonaco, S. Heri, A. Perrenoud, and E. Fischer, “The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli,” *J. Biol. Chem.*, vol. 279, no. 8, pp. 6613–6619, Feb. 2004, doi: 10.1074/jbc.M311657200.

[218] S. Federowicz, D. Kim, A. Ebrahim, J. Lerman, H. Nagarajan, B. Cho, K. Zengler, and B. Palsson, “Determining the Control Circuitry of Redox Metabolism at the Genome-Scale,” *PLOS Genet.*, vol. 10, no. 4, p. e1004264, Apr. 2014, doi: 10.1371/journal.pgen.1004264.

[219] R. Malpica, B. Franco, C. Rodriguez, O. Kwon, and D. Georgellis, “Identification of a quinone-sensitive redox switch in the ArcB sensor kinase,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 36, pp. 13318–13323, Sep. 2004, doi: 10.1073/pnas.0403064101.

[220] K. S. Myers, H. Yan, I. M. Ong, D. Chung, K. Liang, F. Tran, S. Keleş, R. Landick, and P. J. Kiley, “Genome-scale Analysis of Escherichia coli FNR Reveals Complex Features of Transcription Factor Binding,” *PLoS Genet.*, vol. 9, no. 6, p. e1003565, Jun. 2013, doi: 10.1371/journal.pgen.1003565.

[221] P. J. Kiley and H. Beinert, “Oxygen sensing by the global regulator, FNR: the role of the iron-sulfur cluster,” *FEMS Microbiol. Rev.*, vol. 22, no. 5, pp. 341–352, Dec. 1998, doi: 10.1111/j.1574-6976.1998.tb00375.x.

[222] T. M. Ramseier, “Cra and the control of carbon flux via metabolic pathways,” *Res. Microbiol.*, vol. 147, no. 6–7, pp. 489–493, Sep. 1996, doi: 10.1016/0923-2508(96)84003-4.

[223] D. Kim, S. W. Seo, Y. Gao, H. Nam, G. I. Guzman, B.-K. Cho, and B. O. Palsson, “Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP,” *Nucleic Acids Res.*, vol. 46, no. 6, pp. 2901–2917, Apr. 2018, doi: 10.1093/nar/gky069.

[224] K. Kochanowski, B. Volkmer, L. Gerosa, B. R. Haverkorn van Rijsewijk, A. Schmidt, and M. Heinemann, “Functioning of a metabolic flux sensor in Escherichia coli,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 3, pp. 1130–1135, Jan. 2013, doi: 10.1073/pnas.1202582110.

[225] D. Zheng, C. Constantinidou, J. L. Hobman, and S. D. Minchin, “Identification of the CRP regulon using in vitro and in vivo transcriptional profiling,” *Nucleic Acids Res.*, vol. 32, no. 19, pp. 5874–5893, 2004, doi: 10.1093/nar/gkh908.

[226] J. Green, M. R. Stapleton, L. J. Smith, P. J. Artymiuk, C. Kahramanoglou, D. M. Hunt, and R. S.

Buxton, “Cyclic-AMP and bacterial cyclic-AMP receptor proteins revisited: adaptation for different ecological niches,” *Curr. Opin. Microbiol.*, vol. 18, no. 100, pp. 1–7, Apr. 2014, doi: 10.1016/j.mib.2014.01.003.

[227] Y. Miyake, T. Inaba, H. Watanabe, J. Teramoto, K. Yamamoto, and A. Ishihama, “Regulatory roles of pyruvate-sensing two-component system PyrSR (YpdAB) in Escherichia coli K-12,” *FEMS Microbiol. Lett.*, vol. 366, no. 2, Jan. 2019, doi: 10.1093/femsle/fnz009.

[228] K. Rychel, “SBRG/ROS-ALE: Pre-release for initial journal submission.” Zenodo, Dec. 16, 2022. doi: 10.5281/ZENODO.7449004.

[229] E. T. Mohamed, S. Wang, R. M. Lennen, M. J. Herrgård, B. A. Simmons, S. W. Singer, and A. M. Feist, “Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution,” *Microb. Cell Factories*, vol. 16, no. 1, p. 204, Nov. 2017, doi: 10.1186/s12934-017-0819-1.

[230] A. Anand, C. A. Olson, L. Yang, A. V. Sastry, E. Catoiu, K. S. Choudhary, P. V. Phaneuf, T. E. Sandberg, S. Xu, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson, “Pseudogene repair driven by selection pressure applied in experimental evolution,” *Nat. Microbiol.*, vol. 4, no. 3, Art. no. 3, Mar. 2019, doi: 10.1038/s41564-018-0340-2.

[231] S. Chen, T. Huang, Y. Zhou, Y. Han, M. Xu, and J. Gu, “AfterQC: automatic filtering, trimming, error removing and quality control for fastq data,” *BMC Bioinformatics*, vol. 18, no. 3, p. 80, Mar. 2017, doi: 10.1186/s12859-017-1469-3.

[232] D. E. Deatherage and J. E. Barrick, “Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq,” *Methods Mol. Biol. Clifton NJ*, vol. 1151, pp. 165–188, 2014, doi: 10.1007/978-1-4939-0554-6_12.

[233] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biol.*, vol. 10, no. 3, p. R25, Mar. 2009, doi: 10.1186/gb-2009-10-3-r25.

[234] L. Wang, S. Wang, and W. Li, “RSeQC: quality control of RNA-seq experiments,” *Bioinforma. Oxf. Engl.*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012, doi: 10.1093/bioinformatics/bts356.

[235] Y. Liao, G. K. Smyth, and W. Shi, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features,” *Bioinforma. Oxf. Engl.*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.

[236] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinforma. Oxf. Engl.*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[237] H. Latif, R. Szubin, J. Tan, E. Brunk, A. Lechner, K. Zengler, and B. O. Palsson, “A streamlined ribosome profiling protocol for the characterization of microorganisms,” *BioTechniques*, vol. 58, no. 6, pp. 329–332, Jun. 2015, doi: 10.2144/000114302.

[238] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, Art. no. 1, May 2011, doi: 10.14806/ej.17.1.200.

[239] T. Swings, B. Van den Bergh, S. Wuyts, E. Oeyen, K. Voordeckers, K. J. Verstrepen, M. Fauvart, N. Verstraeten, and J. Michiels, “Adaptive tuning of mutation rates allows fast response to lethal

stress in Escherichia coli," *eLife*, vol. 6, p. e22939, May 2017, doi: 10.7554/eLife.22939.

[240] K. Rychel, J. Tan, A. Patel, C. Lamoureux, Y. Hefner, R. Szubin, J. Johnsen, E. T. T. Mohamed, P. V. Phaneuf, A. Anand, C. A. Olson, J. H. Park, A. V. Sastry, L. Yang, A. M. Feist, and B. O. Palsson, "Lab evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance." bioRxiv, p. 2022.12.20.521246, Dec. 23, 2022. doi: 10.1101/2022.12.20.521246.

[241] V. Nguyen, C. Wilson, M. Hoemberger, J. B. Stiller, R. V. Agafonov, S. Kutter, J. English, D. L. Theobald, and D. Kern, "Evolutionary drivers of thermoadaptation in enzyme catalysis," *Science*, vol. 355, no. 6322, pp. 289–294, Jan. 2017, doi: 10.1126/science.aah3717.

[242] C. M. Blatteis, "Fever: pathological or physiological, injurious or beneficial?," *J. Therm. Biol.*, vol. 28, no. 1, pp. 1–13, Jan. 2003, doi: 10.1016/S0306-4565(02)00034-7.

[243] K. Vavitsas, P. D. Glekas, and D. G. Hatzinikolaou, "Synthetic Biology of Thermophiles: Taking Bioengineering to the Extremes?," *Appl. Microbiol.*, vol. 2, no. 1, Art. no. 1, Mar. 2022, doi: 10.3390/applmicrobiol2010011.

[244] T. E. Sandberg, M. Pedersen, R. A. LaCroix, A. Ebrahim, M. Bonde, M. J. Herrgard, B. O. Palsson, M. Sommer, and A. M. Feist, "Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations," *Mol. Biol. Evol.*, vol. 31, no. 10, pp. 2647–2662, Oct. 2014, doi: 10.1093/molbev/msu209.

[245] B. A. Bridges, "Hypermutation in bacteria and other cellular systems.," *Philos. Trans. R. Soc. Lond. Ser. B*, vol. 356, no. 1405, pp. 29–39, Jan. 2001, doi: 10.1098/rstb.2000.0745.

[246] A. Couce, J. R. Guelfo, and J. Blázquez, "Mutational Spectrum Drives the Rise of Mutator Bacteria," *PLoS Genet.*, vol. 9, no. 1, Jan. 2013, doi: 10.1371/journal.pgen.1003167.

[247] P. Modrich, "Mechanisms in E. coli and Human Mismatch Repair (Nobel Lecture)," *Angew. Chem. Int. Ed Engl.*, vol. 55, no. 30, pp. 8490–8501, Jul. 2016, doi: 10.1002/anie.201601412.

[248] M. T. Morita, Y. Tanaka, T. S. Kodama, Y. Kyogoku, H. Yanagi, and T. Yura, "Translational induction of heat shock transcription factor sigma32: evidence for a built-in RNA thermosensor," *Genes Dev.*, vol. 13, no. 6, pp. 655–665, Mar. 1999, doi: 10.1101/gad.13.6.655.

[249] T. Yura, "Regulation of the heat shock response in Escherichia coli: history and perspectives," *Genes Genet. Syst.*, vol. 94, no. 3, pp. 103–108, Jul. 2019, doi: 10.1266/ggs.19-00005.

[250] J. W. Erickson, V. Vaughn, W. A. Walter, F. C. Neidhardt, and C. A. Gross, "Regulation of the promoters and transcripts of rpoH, the Escherichia coli heat shock regulatory gene," *Genes Dev.*, vol. 1, no. 5, pp. 419–432, Jul. 1987, doi: 10.1101/gad.1.5.419.

[251] D. M. Fitzgerald, R. P. Bonocora, and J. T. Wade, "Comprehensive Mapping of the Escherichia coli Flagellar Regulatory Network," *PLoS Genet.*, vol. 10, no. 10, Oct. 2014, doi: 10.1371/journal.pgen.1004649.

[252] K. M. Ottemann and J. F. Miller, "Roles for motility in bacterial-host interactions," *Mol. Microbiol.*, vol. 24, no. 6, pp. 1109–1117, Jun. 1997, doi: 10.1046/j.1365-2958.1997.4281787.x.

[253] H. M. Singer, M. Erhardt, A. M. Steiner, M.-M. Zhang, D. Yoshikami, G. Bulaj, B. M. Olivera, and K. T. Hughes, "Selective Purification of Recombinant Neuroactive Peptides Using the Flagellar Type III Secretion System," *mBio*, vol. 3, no. 3, pp. e00115-12, May 2012, doi: 10.1128/mBio.00115-12.

[254] C. A. Green, N. S. Kamble, E. K. Court, O. J. Bryant, M. G. Hicks, C. Lennon, G. M. Fraser, P. C. Wright, and G. P. Stafford, “Engineering the flagellar type III secretion system: improving capacity for secretion of recombinant protein,” *Microb. Cell Factories*, vol. 18, no. 1, p. 10, Jan. 2019, doi: 10.1186/s12934-019-1058-4.

[255] T. Minamino and K. Namba, “Self-Assembly and Type III Protein Export of the Bacterial Flagellum,” *Microb. Physiol.*, vol. 7, no. 1–2, pp. 5–17, 2004, doi: 10.1159/000077865.

[256] I. Rudenko, B. Ni, T. Glatter, and V. Sourjik, “Inefficient Secretion of Anti-sigma Factor FlgM Inhibits Bacterial Motility at High Temperature,” *iScience*, vol. 16, pp. 145–154, May 2019, doi: 10.1016/j.isci.2019.05.022.

[257] S. Geibel, E. Procko, S. J. Hultgren, D. Baker, and G. Waksman, “Structural and energetic basis of folded-protein transport by the FimD usher,” *Nature*, vol. 496, no. 7444, pp. 243–246, Apr. 2013, doi: 10.1038/nature12007.

[258] J. A. Garnett, V. I. Martínez-Santos, Z. Saldaña, T. Pape, W. Hawthorne, J. Chan, P. J. Simpson, E. Cota, J. L. Puente, J. A. Girón, and S. Matthews, “Structural insights into the biogenesis and biofilm formation by the Escherichia coli common pilus,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 10, pp. 3950–3955, Mar. 2012, doi: 10.1073/pnas.1106733109.

[259] S. Schlegel, E. Rujas, A. J. Ytterberg, R. A. Zubarev, J. Luirink, and J.-W. de Gier, “Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels,” *Microb. Cell Factories*, vol. 12, no. 1, p. 24, Mar. 2013, doi: 10.1186/1475-2859-12-24.

[260] L. H. Gevantman, “Solubility of selected gasses in water”.

[261] K. R. Messner and J. A. Imlay, “The Identification of Primary Sites of Superoxide and Hydrogen Peroxide Formation in the Aerobic Respiratory Chain and Sulfite Reductase Complex of Escherichia coli,” *J. Biol. Chem.*, vol. 274, no. 15, pp. 10119–10128, Apr. 1999, doi: 10.1074/jbc.274.15.10119.

[262] I. Belhadj Slimen, T. Najar, A. Ghram, H. Dabbebi, M. Ben Mrad, and M. Abdrabbah, “Reactive oxygen species, heat stress and oxidative-induced mitochondrial damage. A review,” *Int. J. Hyperthermia*, vol. 30, no. 7, pp. 513–523, Nov. 2014, doi: 10.3109/02656736.2014.971446.

[263] J. W. A. van Beilen and K. J. Hellingwerf, “All Three Endogenous Quinone Species of Escherichia coli Are Involved in Controlling the Activity of the Aerobic/Anaerobic Response Regulator ArcA,” *Front. Microbiol.*, vol. 7, p. 1339, 2016, doi: 10.3389/fmicb.2016.01339.

[264] W. A. Prinz, F. Aslund, A. Holmgren, and J. Beckwith, “The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the Escherichia coli cytoplasm,” *J. Biol. Chem.*, vol. 272, no. 25, pp. 15661–15667, Jun. 1997, doi: 10.1074/jbc.272.25.15661.

[265] D. J. Colquhoun and H. Sørum, “Temperature dependent siderophore production in Vibrio salmonicida,” *Microb. Pathog.*, vol. 31, no. 5, pp. 213–219, Nov. 2001, doi: 10.1006/mpat.2001.0464.

[266] A. T. Smith, R. O. Linkous, N. J. Max, A. E. Sestok, V. A. Szalai, and K. N. Chacón, “The FeoC [4Fe–4S] Cluster Is Redox-Active and Rapidly Oxygen-Sensitive,” *Biochemistry*, vol. 58, no. 49, pp. 4935–4949, Dec. 2019, doi: 10.1021/acs.biochem.9b00745.

[267] G. Jovanovic, P. Mehta, C. McDonald, A. C. Davidson, P. Uzdavinys, L. Ying, and M. Buck,

"The N-terminal amphipathic helices determine regulatory and effector functions of phage shock protein A (PspA) in Escherichia coli," *J. Mol. Biol.*, vol. 426, no. 7, pp. 1498–1511, Apr. 2014, doi: 10.1016/j.jmb.2013.12.016.

[268] C. Marotz, A. Amir, G. Humphrey, J. Gaffney, G. Gogul, and R. Knight, "DNA extraction for streamlined metagenomics of diverse environmental samples," *BioTechniques*, vol. 62, no. 6, pp. 290–293, Jun. 2017, doi: 10.2144/000114559.

[269] T. C. Glenn, R. A. Nilsen, T. J. Kieran, J. G. Sanders, N. J. Bayona-Vásquez, J. W. Finger, T. W. Pierson, K. E. Bentley, S. L. Hoffberg, S. Louha, F. J. G.-D. Leon, M. A. del R. Portilla, K. D. Reed, J. L. Anderson, J. K. Meece, S. E. Aggrey, R. Rekaya, M. Alabady, M. Belanger, K. Winker, and B. C. Faircloth, "Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext)," *PeerJ*, vol. 7, p. e7755, Oct. 2019, doi: 10.7717/peerj.7755.

[270] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

[271] E. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," 1960.

Appendix A. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome

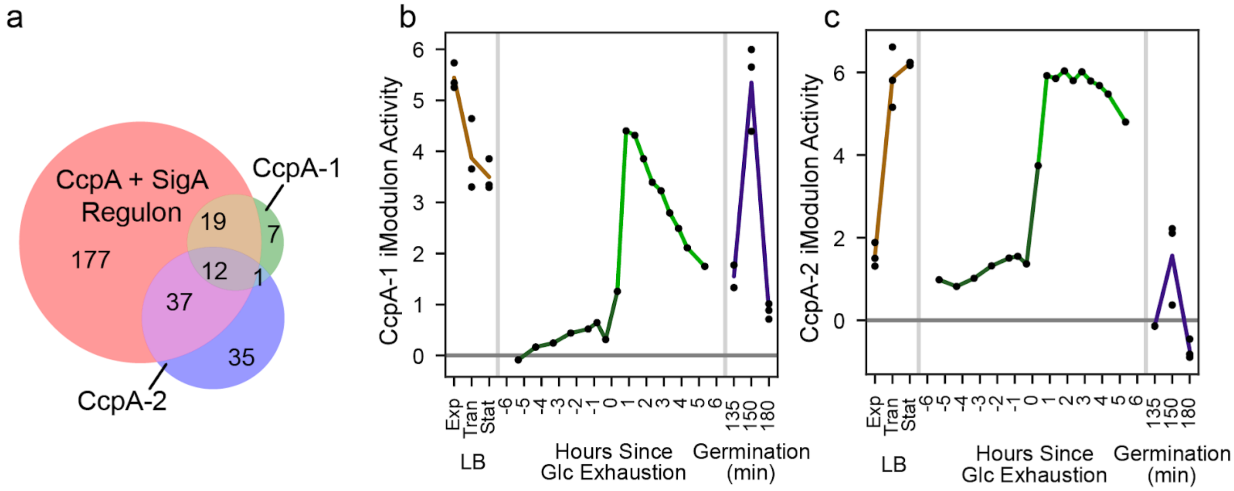


Figure A.1: The CcpA iModulons. CcpA-1 contains mostly sugar metabolism enzymes (ribose, sucrose, mannose, trehalose, lichenan, etc.) while CcpA-2 contains a mix of genes including those for inositol consumption, tricarboxylic acid permeability, and acetyl-CoA utilization. **(a)** Venn diagram of gene membership for these iModulons and their matched regulon. **(b-c)** Activity of CcpA iModulons for three experiments: growth in LB media ("Exp", "Tran" and "Stat" refer to exponential, transition, and stationary phase, respectively), glucose (Glc) exhaustion, and germination. Dots indicate individual samples and lines pass through means. **(b)** CcpA-1 is active during exponential growth and germination, but declines in stationary phase and during glucose exhaustion. **(c)** CcpA-2 is active during stationary phase and throughout the first 5 hours of glucose exhaustion, and comparatively less active during germination.

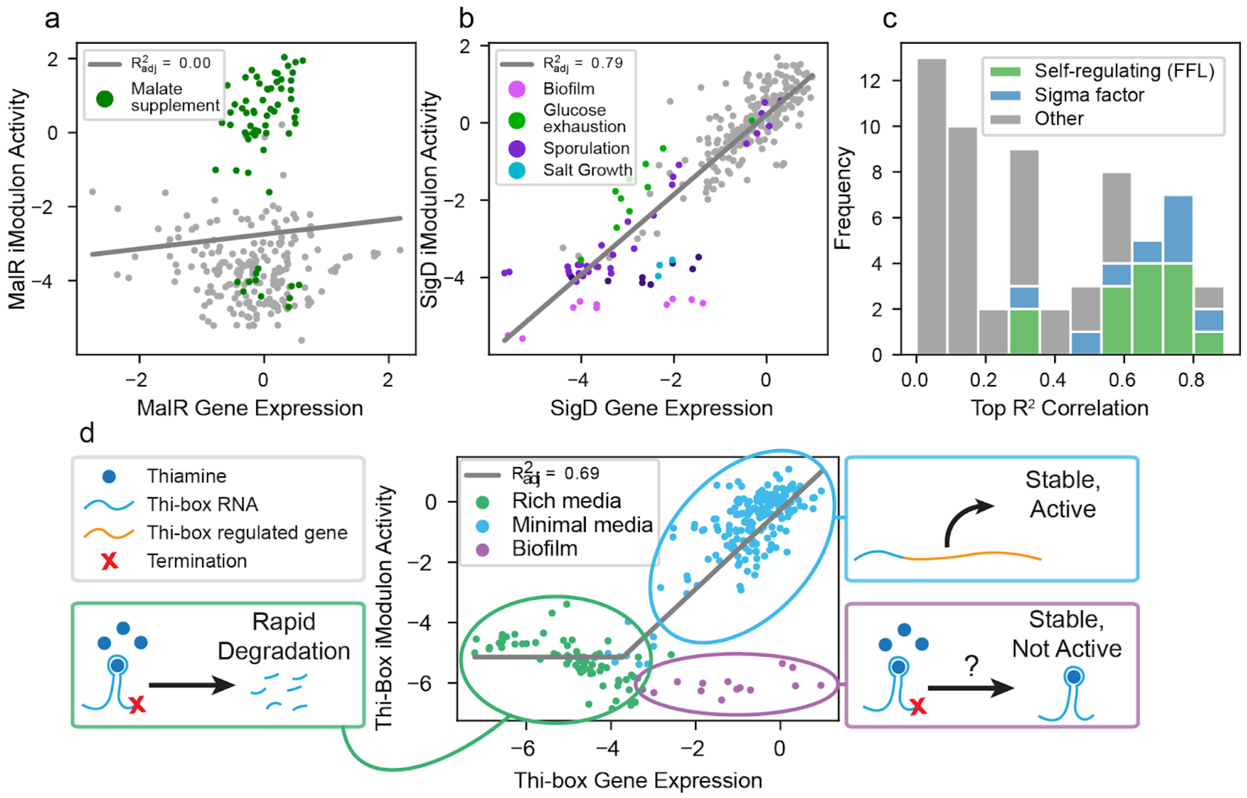
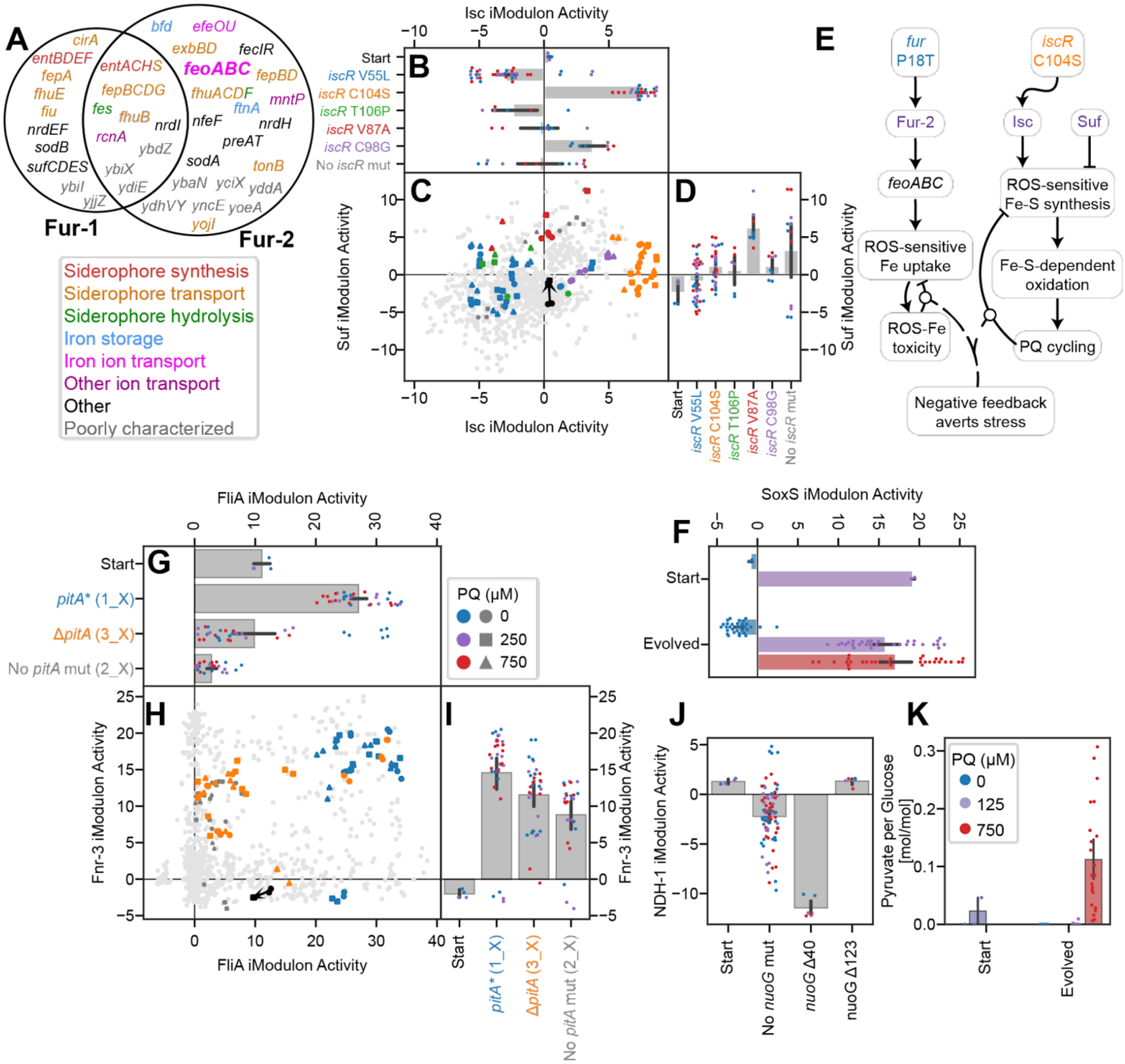


Figure A.2: Correlation between iModulon activity and regulator expression. Each plot contains an iModulon's transcriptional regulator expression on the x axis and the corresponding activity level on the y axis. **(a)** MalR is a typical transcription factor that responds to kinase activity downstream of malate binding. The activity is therefore not correlated with regulator expression but is increased with malate supplementation. **(b)** SigD is the sigma factor governing motility, which is regulated at the transcriptional level, so a correlation between activity and expression is observed. Four experiments that exhibit low activity are highlighted; their activity matches expectations from literature. **(c)** Histogram of correlations for all iModulons with a known regulator. Self-regulating iModulons are those that contain a TF which also regulates them (FFL: Feed-Forward Loop). Higher correlations are observed for FFL and sigma factor-regulated iModulons. **(d)** This Thi-box transcript precedes *thiC*, and the other 4 Thi-box transcripts exhibit similar patterns. A broken line was used for this regression. When activity is low, the expressed RNA contains only the short Thi-box sequence, which appears to be degraded quickly in the rich media condition (flasks) but not in the biofilm condition. The Thi-box RNA appears to be stable in biofilms for an unknown reason.

Appendix B. Laboratory Evolution, Transcriptomics, and Modeling Reveal Mechanisms of Paraquat Tolerance

Figure B.1: Additional insights from mutational, iModulon, and metabolic analysis. Bars indicate mean ± 95% confidence interval. **(A)**Venn diagram of the Fur-1 and Fur-2 iModulon genes, color coded by function. Ion transport and storage systems, which may be advantageous under ROS conditions, are enriched in Fur-2. **(B-D, H-J)** Scatter plots of iModulon activities with bar plots sharing axes. Light gray dots indicate other samples from PRECISE-1K. In (C) and (H), samples are colored by relevant mutations, and shapes indicate PQ concentrations according to the legends. A black arrow connects the starting strain samples between 0 and 250 μM PQ. In bar plots, point colors indicate PQ concentrations and label colors match with the scatter plots. **(B-D)** Suf and Isc iModulon activities, which are both regulated by IscR and encode distinct Fe-S cluster synthesis mechanisms (Suf is more robust to stress compared to Isc). **(E)** Knowledge graph linking two key TF mutations through their iModulons to negative feedback which averts stress. **(F)** SoxS iModulon activity is correlated with PQ in both starting and evolved strains (Pearson R = 0.72, p = $5.5*10^{-15}$). **(G-I)** FliA and Fnr-3 iModulon activities by *pitA* mutation, showing an unexpected upregulation in the case of the frameshift *pitA**, but not in the case of the *pitA* deletion. **(J)** NDH-1 iModulon activities. The NDH-1 iModulon consists of genes (*nuoGHIJKLMN*) that are controlled by ArcA and Fnr and are all downstream of the *nuoG* Δ40 mutation, which may create a terminator sequence. **(K)** Pyruvate production rates from exometabolomic characterizations of evolved strains. Note that the starting strain was characterized at 0 and 125 μM PQ (due to no growth at higher PQ), whereas the evolved strains were characterized at 0, 250, and 750 μM PQ. Pyruvate is secreted at high PQ levels, particularly by evolved strains which have downregulated PDH and the TCA cycle.


A
cirA
entBDEF
fepA
fhuE
fiu
nrdEF
sodB
sufCDES
ybiI
yjjZ
entACHS
fepBCDG
fes
fhuB
nrdI
rcnA
ybdZ
ybiX
ydiE
bfd
efeOU
exbBD
fecIR
feoABC
fepBD
fhuACDF
mntP
ftnA
nfeF
nrdH
preAT
sodA
tonB
ybaN
yciX
yddA
ydhVY
yncE
yoeA
yojI
Fur-1
Fur-2
Siderophore synthesis
Siderophore transport
Siderophore hydrolysis
Iron storage
Iron ion transport
Other ion transport
Other
Poorly characterized
Isc iModulon Activity
−10
−5
0
5
B
Start
iscR V55L
iscR C104S
iscR T106P
iscR V87A
iscR C98G
No iscR mut
C
Suf iModulon Activity
10
5
0
−5
−10
Isc iModulon Activity
D
Suf iModulon Activity
E
fur P18T
iscR C104S
Fur-2
Isc
Suf
feoABC
ROS-sensitive Fe-S synthesis
ROS-sensitive Fe uptake
Fe-S-dependent oxidation
ROS-Fe toxicity
PQ cycling
Negative feedback averts stress
FliA iModulon Activity
0
10
20
30
40
G
Start
pitA* (1_X)
ΔpitA (3_X)
No pitA mut (2_X)
PQ (μM)
0
250
750
H
Fnr-3 iModulon Activity
25
20
15
10
5
0
−5
FliA iModulon Activity
I
Fnr-3 iModulon Activity
Start
pitA* (1_X)
ΔpitA (3_X)
No pitA mut (2_X)
SoxS iModulon Activity
−5
0
5
10
15
20
25
F
Start
Evolved
J
NDH-1 iModulon Activity
5
0
−5
−10
Start
No nuoG mut
nuoG Δ40
nuoG Δ123
K
Pyruvate per Glucose [mol/mol]
0.3
0.2
0.1
0.0
PQ (μM)
0
125
750
Start
Evolved

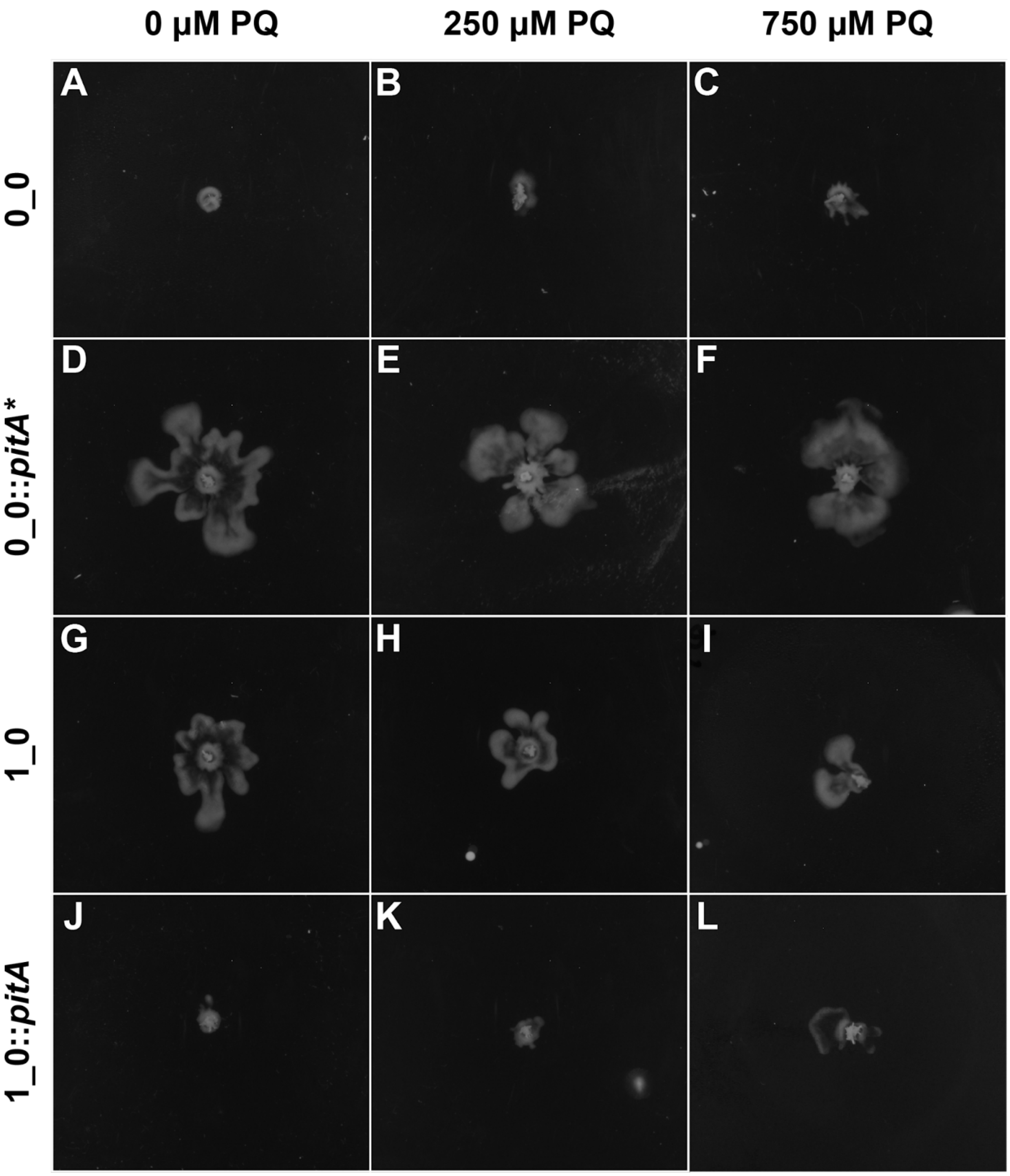


Figure B.2: Swarming assays of *pitA* mutants. Cells were plated on agar in tryptone broth with glucose and the PQ concentration shown in the column headers. They were allowed to swarm for one day prior to image capture. The *pitA* mutant strains 0_0::*pitA** **(D-F)** and 1_0 **(G-I)** swarmed, while wild type *pitA* strains **(A-C; J-L)** did not. Panels A and D are shown in **Figure 4.5K**.

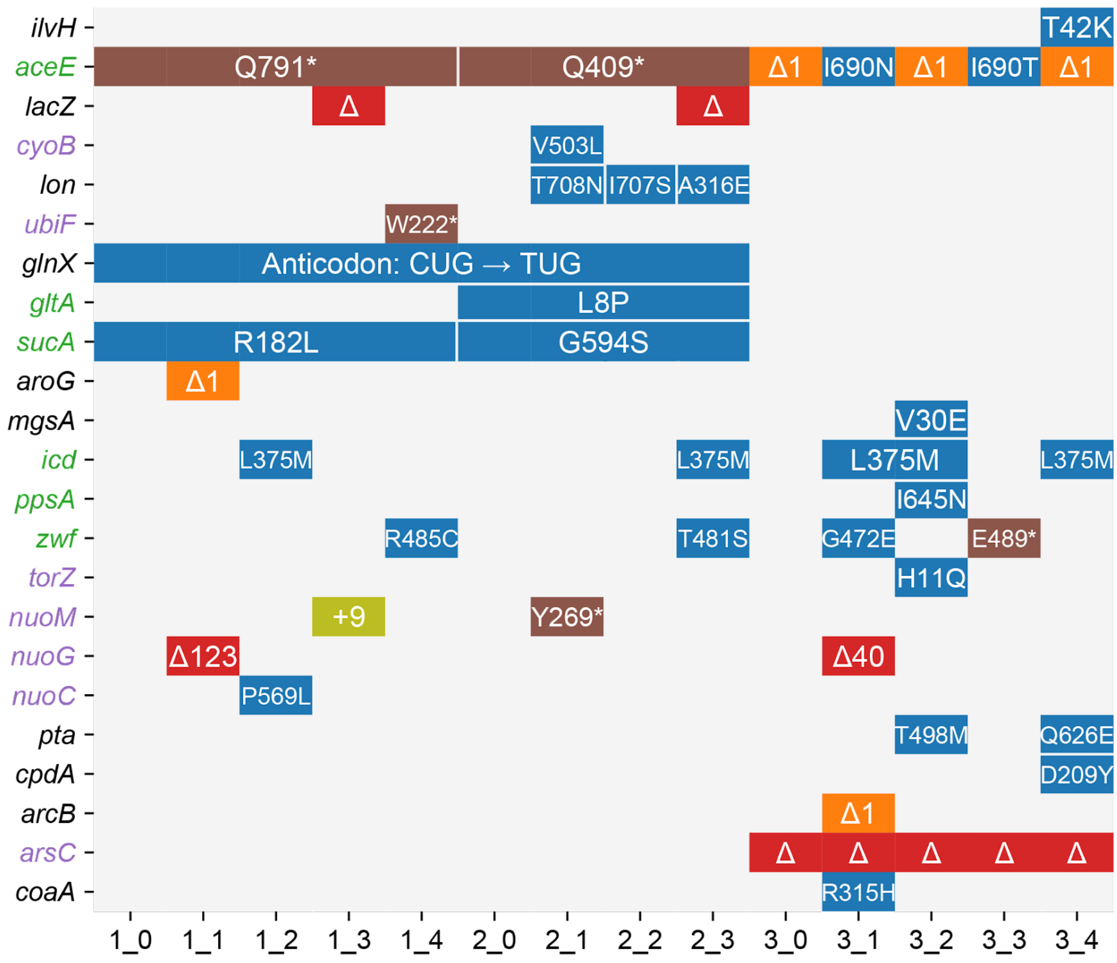


Figure B.3: Mutations in genes relevant to metabolism. Colored blocks share a mutation in the given strain and gene (blue: missense SNP; brown: nonsense SNP; orange: frameshift deletion less than 3 bp; red: large deletion affecting gene; olive: insertion that does not cause a frameshift). Gene names are colored by type (green: central carbon metabolic enzyme; purple: redox enzyme; black: other gene relevant to metabolism). Silent mutations, the e14 deletion, and promoter mutations are omitted.

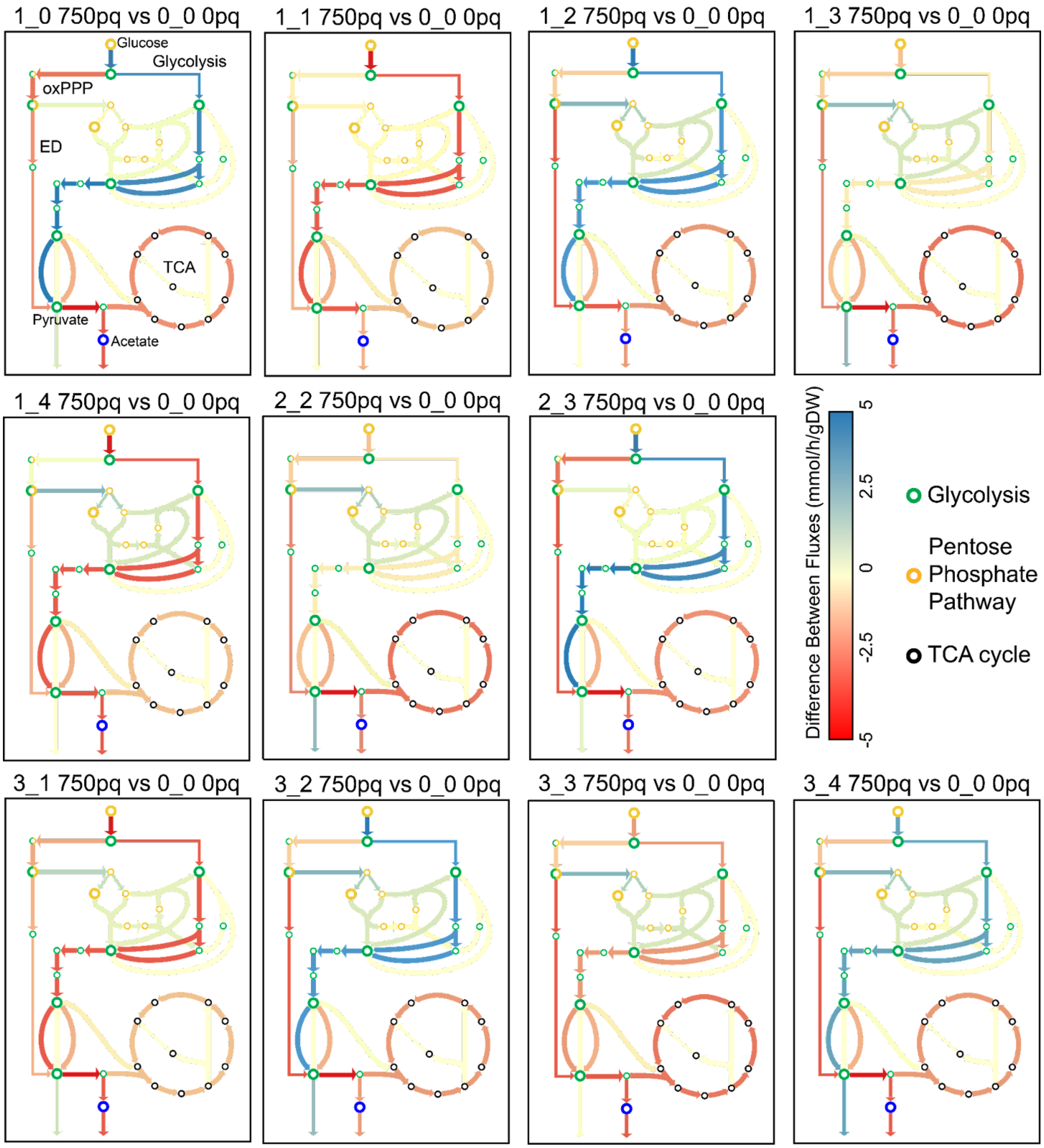


Figure B.4: The constrained OxidizeME model predicts the flux distribution change in central metabolism after evolution. Flux distribution changes from specific OxidizeME models, constrained by RNAseq, growth, and glucose uptake data. TCA cycle flux always decreases after evolution (**Figure 4.7D**), and glycolytic flux varies with glucose uptake rate. Note that glucose uptake increases in evolved strains relative to the stressed starting strain, but some strains have more or less glucose uptake relative to the unstressed starting strain.

Appendix C. Laboratory evolution reveals transcriptional mechanisms underlying thermal adaptation of *Escherichia coli*

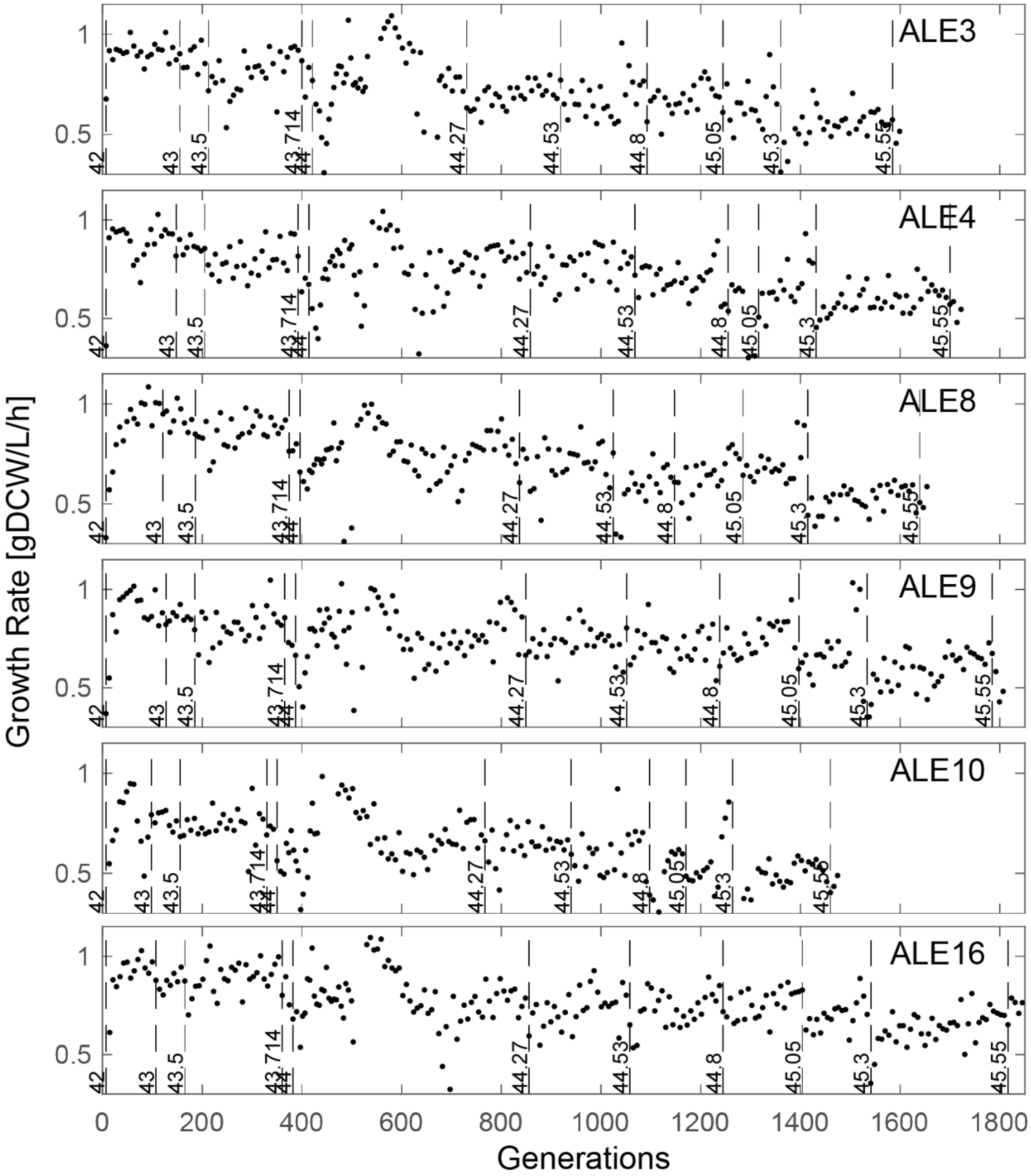


Figure C.1: Growth rates and temperatures for each flask in ALE. Dotted vertical lines indicate the flasks at which the temperature was increased. Generation numbers are estimated from the growth rate and elapsed time of each flask.