Title

Understanding transcriptional regulatory mechanisms through data science and modeling

Permalink

https://escholarship.org/uc/item/7nn3074j

Author

Dalldorf, Christopher

Publication Date

2024

UNIVERSITY OF CALIFORNIA SAN DIEGO

Understanding transcriptional regulatory mechanisms through data science and modeling

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

in

Bioengineering

by

Christopher Gilbert Dalldorf

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Jim Kadonaga
Professor Rob Knight
Professor Sergey Kryazhimskiy
Professor Joseph Pogliano

2024

The dissertation of Christopher Gilbert Dalldorf is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To Family, Friends, and Good Food.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost I want to thank my family. Talking with my mom every Sunday is how I've wrapped up every week since I left for college and she was always the first person I wanted to tell when a paper was accepted or my data looked promising. I greatly appreciate all of the time and care you've spent just listening to my everyday life. I'd like to thank my dad for being exactly who he is, the most positive, happy, and loving person I've ever had the pleasure to know. Your advice is always way more helpful than I think you know it is. To my sisters, I love each of you for different but much the same reason. Katie has always been my big sister and role model who I aspire to work as hard as, Delaney can always cheer me up on a sad day and keep me feeling positive, and I can't thank Sophie enough for helping me be more open with myself and others. I can't forget my grandma, who is a delight to talk to and I can always call to give me exactly the advice that I want. I've never questioned how much you all love me and that's given me tremendous strength throughout my life.

I'd also like to thank my friends here in San Diego who have made San Diego home. Y'all have helped me through a lot and I greatly appreciate both the shoulders to lean on when needed and exorbitant time spent managing fictional kingdoms, both medieval and space-faring. I love y'all Tony, Jeff, Dom, Jon, Laura, Lucas, Marc, Lina, and I'm sure others I'm forgetting. I wouldn't have been able to do this without the support. I also need to thank my friends who don't live here who have stayed in touch, whether with Will through bad movie franchises, sharing good music with Hunter, fictional roleplaying with my DnD group, or way too many hours optimizing fictional factories with Matthew.

Of course I need to thank my lab mates, both for being sage sources of technical and intellectual advice and for making the lab a brighter place. The pandemic interrupted a lot

of this, but it makes me feel warm to see people solving scavenged jigsaw puzzles, having tiny basketball shooting competitions, and debating if pretzels and grapes taste like chicken. I also need to thank some past lab mates, namely Patrick Phaneuf and Anand Sastry for mentoring me when I first joined the lab. Patrick stopping by my desk every week and giving me a few to-do's as well as sharing his personal passion for the field led me to switch from a master's student to a PhD.

I need to thank Dr. Palsson for being willing to mentor me through the painful, meticulous, and sometimes rewarding process of writing papers. I believe I ended up at 22 drafts for the first paper if I counted correctly and I'm sure you grew as tired as I did of reading them, so I greatly appreciate your feedback. I also need to thank Daniel Zielinski for both being a large part of nearly all of my research and for answering my constant questions, both academic and occasionally about life. I'm certain I would not have graduated without you.

I also need to thank a bunch of people from throughout my life. The Griffins for being my second family. Jon Schner, Linda Sudnik, and Mike Gale for saving my life all those years ago. Dr. Buck for assuring me I could live a normal life afterwards. My scoutmaster Vance Barron for starting my love and appreciation of nature. Mr. Register for teaching me how awesome science can be. Professor Simon Shepherd for getting me interested in applying computer science to academia. I'm sure I'm forgetting others, so if you aren't listed here I probably just forgot in the madness that is wrapping up a PhD.

Chapter 2 in part is a reprint of material published in:

- **C Dalldorf**, K Rychel, R Szubin, Y Hefner, A Patel, DC Zielinski, BO Palsson. 2024. "The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions" *mSystems*, 10.1128/msystems.00305-24. The dissertation author was the primary author.

Chapter 3 in part is a reprint of material submitted for publication in *Proceedings of the National Academy of Sciences (PNAS)*:

- **C Dalldorf**, Y Hefner, R Szubin, J Johnsen, E Mohamed, G Li, J Krishnan, AM Feist, BO Palsson, DC Zielinski. 2024. "Diversity of transcription regulatory adaptation in *E. coli*" The dissertation author was the primary author.

Chapter 4 in part is a reprint of material submitted to *bioRxiv*:

- **C Dalldorf**, G Hughes, G Li, BO Palsson, DC Zielinski. 2024. "Data-driven modeling of bacterial transcriptional regulation" The dissertation author was the primary author.

VITA

2016	Bachelor of Arts in Engineering Sciences, Dartmouth College
2016	Bachelor of Engineering in Biomedical Engineering, Dartmouth College
2020	Master of Science in Bioengineering, University of California San Diego
2024	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

C Dalldorf, K Rychel, R Szubin, Y Hefner, A Patel, DC Zielinski, BO Palsson. 2024. ”The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions” *mSystems*, 10.1128/msystems.00305-24.

PV Phaneuf, DC Zielinski, JT Yurkovich, J Johnsen, R Szubin, L Yang, SH Kim, S Schulz, M Wu, **C Dalldorf**, E Ozdemir, RM Lennen, BO Palsson, AM Feist. 2021. ”Escherichia coli Data-Driven Strain Design Using Aggregated Adaptive Laboratory Evolution Mutational Data” *ACS Synthetic Biology*, 10.1021/acssynbio.1c00337.

A Rajput, Y Seif, KS Choudhary, **C Dalldorf**, S Poudel, JM Monk, BO Palsson. 2021. ”Pangenome analytics reveal two-component systems as conserved targets in ESKAPEE pathogens” *mSystems*, 10.1128/mSystems.00981-20.

S Chowdhury, DC Zielinski, **C Dalldorf**, , JV Rodrigues, BO Palsson, and EI Shakhnovich. 2021. ”Empowering drug off-target discovery with metabolic and structural analysis” *Nat Commun*, 10.1038/s41467-023-38859-x

AJ Martin, D Jenkins, R Zhang, **C Dalldorf**, VC Douglas. 2017. ”Consistent corrosion progression preceding gross taper failure in THR of a single design” *Trans ORS*.

ABSTRACT OF THE DISSERTATION

Understanding transcriptional regulatory mechanisms through data science and modeling

by

Christopher Gilbert Dalldorf

Doctor of Philosophy in Bioengineering

University of California San Diego, 2024

Professor Bernhard Ø. Palsson, Chair

Transcriptional gene regulation is a primary mechanism that *Escherichia coli* uses to best adapt to its current environment. RNA-sequencing data in particular provides us with the ability to more clearly examine the different states of the transcriptome and thus study transcriptional regulation. The cost to generate RNA-sequencing datasets has dramatically decreased over the last two decades which, in conjunction with FAIR data principles, has subsequently led to a large increase in the amount of publicly available RNA-sequencing data. In addition, centralized public databases of both these large datasets and biological knowledge have become increasingly large and accessible. Both developing and deploying new analytical techniques are necessary in order to best derive actionable insights from this large increase in scale for biological research. In this thesis, we first apply these principles to studying RNAP mutations and their ability to shift a transcriptome to favor growth over stress functions. This tradeoff between fear and greed can be seen in nearly all RNAP mutations and also across numerous bacterial species. Next we investigate the plasticity of transcriptional regulation by removing selected transcription factors and evolving strains, finding some knockouts recover growth without significant adaptation while others require convergent mutations to regulatory elements which restore the expression of highly growth-important genes. Finally, we build a model for the transcriptional regulatory network using iModulons as a measure of regulator activity. This model is able to accurately predict metabolite concentrations as well as infer biological constants about transcription factor binding. Altogether, these projects help advance our understanding of the mechanisms underlying transcriptional regulation through the utilization of data science.

Chapter 1

Transcriptional regulation as part of the central dogma of molecular biology

The central dogma of molecular biology, that of the pathway from DNA to RNA to proteins, is at the core of all life. The conversion of DNA to RNA is called transcription and involves a large diversity of components. Transcription factors (TFs) and other regulators bind to DNA to recruit or obstruct RNA polymerase (RNAP) binding, which if correctly bound transcribes the genes into RNA. Which TFs are active [1], the concentrations of metabolites in the cell [2], the growth phase of the cell [3], the number of ribosomes available [4], the various possible conditions of the medium (pH, temperature, carbon source, etc.) [5], and many other factors can modify the expression state of the cell by acting on these regulators and RNAP directly. In order to fully understand and model this process, both thorough genetic annotation and large amounts of data are necessary.

RNA-sequencing (RNA-seq) measures the concentration of mRNA in a cell and the amount of RNA-seq data available has dramatically increased over the last decade, largely due to a substantial reduction in cost. In 2012, Illumina RNA-seq cost $275 per sample [6] which by 2023 dropped to $210 [7]. When adjusted for inflation ($275 becomes $365), this is a reduction of 42% in cost per sample. The National Center for Biotechnology Information (NCBI) began the Sequence Read Archive in 2007 which collects both RNA-seq and ChIP-Seq datasets [8]. As of 2024, there are over 10,000 publicly available *E. coli* transcriptomes in this dataset, a number that has more than doubled in the last five years [8].

In addition to this massive increase in public data, the concatenation of biological knowledge to public databases has vastly improved access to the highly expensive and invaluable information created by countless hours of productive microbiology research. EcoCyc, a bioinformatics database containing genome annotations about *E. coli*, has grown from 2,537 citations in 2012 to 3,791 in 2022 [9]. RegulonDB, a knowledge base of transcriptional regulation, was first released in 1998 with 533 regulatory interactions [10] and has since grown to 6,110 in 2023 [11]. Numerous additional highly valuable public databases exist, such as UniProt [12], KEGG [13], RCSB [14], and STRING [15] among others. Additional pertinent databases to this thesis are ALEdb [16], a database of mutations derived from evolution experiments, and iModulonDB [5], a source for data-driven pseudo-regulons named iModulons.

These open sources of biological information enable a dramatic increase in the discovery potential of computational biology. FAIR data principles (Findable, Accessible, Interoperable, and Reusable) increase scientific cooperativity and enable the research of hitherto impossible to study questions [17]. Global studies on the transcriptomics of entire species are now feasible with sufficient computing power and bioinformatics knowledge. This large increase in data size creates a new demand for novel analysis techniques to take advantage of this publicly available data and create widely interpretable results. Additionally, this background of data provides a plethora of invaluable comparisons for any future studies if correctly utilized.

1.1 The primary steps in gene regulation

Like most research, the first earnest effort into studying genetic regulation stemmed from work and interest with a well characterized element. For gene regulation, this was the *lac* operon, which contains the genes necessary for lactose uptake and utilization [18]. The *lac* operon serves as the guiding example for the organization of regulatory elements of gene operons. Upstream of the *lac* operon transcription start site are binding sites for *lacI* and *crp* which repress and promote, respectively, the expression of said operon [19]. *LacI* and *crp* then recruit or prevent the recruitment of RNAP to the operon [20]. If successfully recruited, RNAP then transcribes the operon's genes into mRNA which is later passed to a ribosome to be translated into proteins.

The process outlined above, however, is simplistic and avoids some of the additional complexities of transcriptional regulation. Sigma factors also bind to RNAP and have large roles in changing global gene expression by modifying the binding properties of RNAP to DNA [21]. These sigma factors can modulate the expression of hundreds of genes, often genes specifically related to a phenotypic response such as stationary phase, stress response, sporulation, and virulence [21]. Although not largely explored in this thesis, other additional transcriptional regulatory mechanisms exist such as antitermination complexes which prevent early transcriptional pausing [22], connections to translation via interactions with tRNAs [23], RNAP availability and dynamics [24], and various interactions with ncRNA [25] among many others. In addition to these, gene regulation may also occur post-transcriptionally [26] or even post-translationally [27].

Taken together, all of these various steps in gene regulation work towards the primary goal of best adapting the available proteome of the cell to the growth condition [27–29]. Cells can make large adjustments to their regulatory networks and growth phenotypes through single allele changes [30]. A more complete knowledge and model of gene regulation and the adaptability of the transcriptional regulatory network (TRN) can enable better strain engineering [31] and improved understanding of infectious strains [32].

1.2 Increase in scale of data, of gene annotations, and of characterized TF binding sites

As mentioned earlier, both transcriptional data and knowledge about the TRN has greatly expanded over the past two decades, in large part due to central databases such as NCBI and regulonDB. Only recently has genome- or even multi-genome wide transcriptional research become feasible thanks to this influx of information.

While it is not obvious what regulatory connections are omitted from such a database, 53% of E. coli's genes now have at least one annotated regulator with confirmed or strong confidence in regulonDB [33]. These genes account for 67% of the variance seen in PRECISE1K, a large compendium of gene expression data [34]. On EcoCyc, 58% of genes are well-characterized. These genes explain 64% of the variance in PRECISE1K. 80% of genes in EcoCyc are at least partially characterized which explain 83% of PREICSE1K's variance. Altogether, the knowledge surrounding *E.coli* has grown to a critical mass enabling both larger scale analyses and thorough verification of detailed studies.

1.3 Application of machine learning to RNA-seq

This plethora of data and biological knowledge creates a new opportunity and challenge of scale. Improved data analytical techniques now are in high demand in order to obtain meaningful and applicable biological results from what is, in the case of large-scale transcriptomics research, sometimes millions of individual data points [34]. For transcriptomics over the last fifteen years, the state of the art for analysis has largely been differential gene expression [35]. This process involves finding statistically differentially expressed genes based on expression fold-changes and the normal variance of the genes.

While this can be highly informative for experiments with specific and direct hypotheses, on the majority of genome-scale analyses this approach generates far more potential genes of interest than can be reasonably analyzed. Differential gene expression also relies upon direct comparison between two samples, making multi-dimensional comparisons infeasible. In order to enable higher-dimensional comparisons, numerous data analytical tools have been developed. Many methods for large-scale transcriptomics comparisons have been tested [36], but independent component analysis (ICA) has arisen as a highly informative and non-biased algorithm to modularize transcriptional data into comprehensible subunits called iModulons [37].

The process of generating iModulons using ICA is well explained at https://imodulondb.org/about.html, but can be best understood as a blind-source separation of regulatory signals. ICA is able to separate the gene expression matrix (X) of numerous samples into an individual signal matrix (M) and an activity matrix (A). M contains sets of genes called iModulons that are co-regulated and highly correlated across thousands of experiments while A contains the activity of these various iModulons across said experiments. These iModulons often have high overlap with known regulons of regulators such as sigma factors and transcription factors. This technique is invaluable for computing the regulatory activity of groups of genes across thousands of samples and also for contextualizing new experiments by providing a compendium of comparison points.

1.4 Thesis outline

The three primary aims of my thesis all use data analytical techniques to connect together disparate data sources in order to further expand our knowledge of transcriptional regulation in *E. coli*. The first is a study of RNAP mutations that are commonly found in adaptive laboratory evolutions. Twelve RNAP mutations were reintroduced into *E. coli* where they created large adjustments in the transcriptome, generally to favor faster growth by reducing the expression of stress-related genes and upregulating the expression of ribosomal subunits. The second is a collection of transcription factor knockout evolutions that were carried out in order to better understand the adaptive mechanisms available to the TRN. 11 TF's were knocked out and the resulting strains were evolved, sequenced, expression profiled, and tested for substrate readiness. Some TF KO's recovered using convergent mutations to their own regulatory networks while many TF KO's recovered without TF-specific adaptations. The third aim is the development of a transcriptional model which uses iModulons as a proxy for regulator activity in order to create a data-informed model of gene regulation. This process can accurately predict metabolite concentrations and TF binding constants using a constrained mechanical mathematical model.

Chapter 2

The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions

Fast growth phenotypes are achieved through optimal transcriptomic allocation, in which cells must balance tradeoffs in resource allocation between diverse functions. One such balance between stress readiness and unbridled growth in *E. coli* has been termed the fear versus greed (f/g) tradeoff [37]. Two specific RNA polymerase (RNAP) mutations observed in adaptation to fast growth have been previously shown to affect the f/g tradeoff [30], suggesting that genetic adaptations may be primed to control f/g resource allocation. Here we conduct a greatly expanded study of the genetic control of the f/g tradeoff across diverse conditions. We introduced twelve RNA polymerase (RNAP) mutations commonly acquired during adaptive laboratory evolution (ALE) and obtained expression profiles of each. We found that these single RNAP mutation strains resulted in large shifts in the f/g tradeoff primarily in the RpoS regulon and ribosomal genes, likely through modifying RNAP-DNA interactions. Two of these mutations additionally caused condition-specific transcriptional adaptations. While this tradeoff was previously characterized by the RpoS regulon and ribosomal expression, we find that the GAD regulon plays an important role in stress readiness and ppGpp in translation activity, expanding the scope of the tradeoff. A phylogenetic analysis found the greed-related genes of the tradeoff present in numerous bacterial species. The results suggest that the f/g tradeoff represents a general principle of transcriptome allocation in bacteria where small genetic changes can result in large phenotypic adaptations to growth conditions.

2.1 Background

Maintaining optimal fitness in microorganisms requires navigating tradeoffs in resource allocation [29] due to dependencies between growth and expression [38–40]. High growth rate expression states have been shown to downregulate stress response genes ("fearful" genes) and upregulate ribosomal genes ("greedy" genes) [29]. Furthermore, this tradeoff has been well documented in adaptive laboratory evolution (ALE) experiments [37, 41–43].

We have recently shown that transcriptional shifts of the *E. coli* transcriptome can be viewed through the use of a novel transcriptomic analysis method which uses independent component analysis on large-scale expression databases to define sets of genes that are independently modulated, forming data-driven regulons termed iModulons [37]. Through this analysis, we have identified that the fear versus greed (f/g) tradeoff is characterized by the strong negative correlation between the activity levels of the RpoS (fear) and Translation (greed) iModulons. The f/g tradeoff involves an upregulation of ribosomal genes (greed represented by the Translation iModulon) that often are the limiting factor for increasing growth rate [44] and a concurrent downregulation of stress-related genes (fear represented by the RpoS iModulon). While these iModulons and genes do not encompass all potential growth- and stress-related iModulons and genes within *E. coli*, they are unique in that they follow this tradeoff across a wide variety of conditions.

In addition to the two primary f/g iModulons, the GadX iModulon is also involved in the fear response while the ppGpp iModulon adds another dimension to greed. The Translation iModulon primarily consists of ribosomal subunits, the RpoS iModulon contains the general stress response sigma factor RpoS's regulon, the GadX iModulon is related to acid stress, and the ppGpp iModulon is composed of genes involved in protein translation rates and the stringent response. The tradeoff between these sets of iModulons involves competition between the housekeeping and stress sigma factors (RpoD and RpoS), binding of ppGpp and DksA to RNAP which modifies which genes RNAP transcribes, and other regulatory mechanisms [28, 45, 46]. Many of these mechanisms directly involve RNA polymerase (RNAP) whose availability, along with sigma factor competition, has been previously connected to said tradeoff [47].

RNAP mutations have been shown to drive the f/g tradeoff towards faster growth and RNAP is one of the most common mutation targets during ALEs [16]. In a detailed study of two RNAP mutations found in the catalytic center, it was hypothesized these RNAP mutations adjust the tradeoff towards greed by destabilizing the rpoB-rpoC interface, thus affecting the binding of ppGpp to RNAP [30]. While many ALE mutations cluster in the catalytic center of RNAP, there are numerous other RNAP mutations found in ALE endpoint strains. These mutations can be found near regulator binding sites, regions known to be related to antibiotic resistance, important structural elements such as the flap domain and trigger loop, and in regions with no clear annotations [48–56]. Convergent RNAP mutations have been found in specific environmental adaptation experiments [57–60] often leading to the assumption that RNAP mutations reflect media adaptations, missing their underlying role in the f/g tradeoff. Despite being highly common evolutionary adaptations, the effect of these mutations is largely unknown.

Here, we sought to expand our knowledge of these RNAP mutations and the f/g tradeoff through a multi-scale study incorporating FAIR (Findable, Accessible, Interoperable, Reusable) data principles by using previously generated data and creating new easily accessible data [17]. We first gathered the existing data on RNAP mutations and selected twelve mutations to address the shortcomings of said existing data. We then introduced the twelve RNAP mutations and used computer simulations to infer how these mutations destabilize RNAP. We then obtained transcriptomes in various experimental conditions and used iModulon analysis to demonstrate that, despite structurally distinct locations, these mutations nearly universally downregulate stress-related genes and upregulate growth-related genes (see A.1) in addition to some condition-specific adaptations. We explored additional dimensionality of the tradeoff involving the ppGpp and GadX iModulons. Finally, we compared the transcriptomes of various species to find that that f/g tradeoff is widely found across phylogeny. Thus, our multi-scale study elucidated key features of a central transcriptomic tradeoff between fear and greed in which cells that favor faster growth face the cost of diminished responsiveness to stresses [30] and proposed that it is a general principle in microbiology.

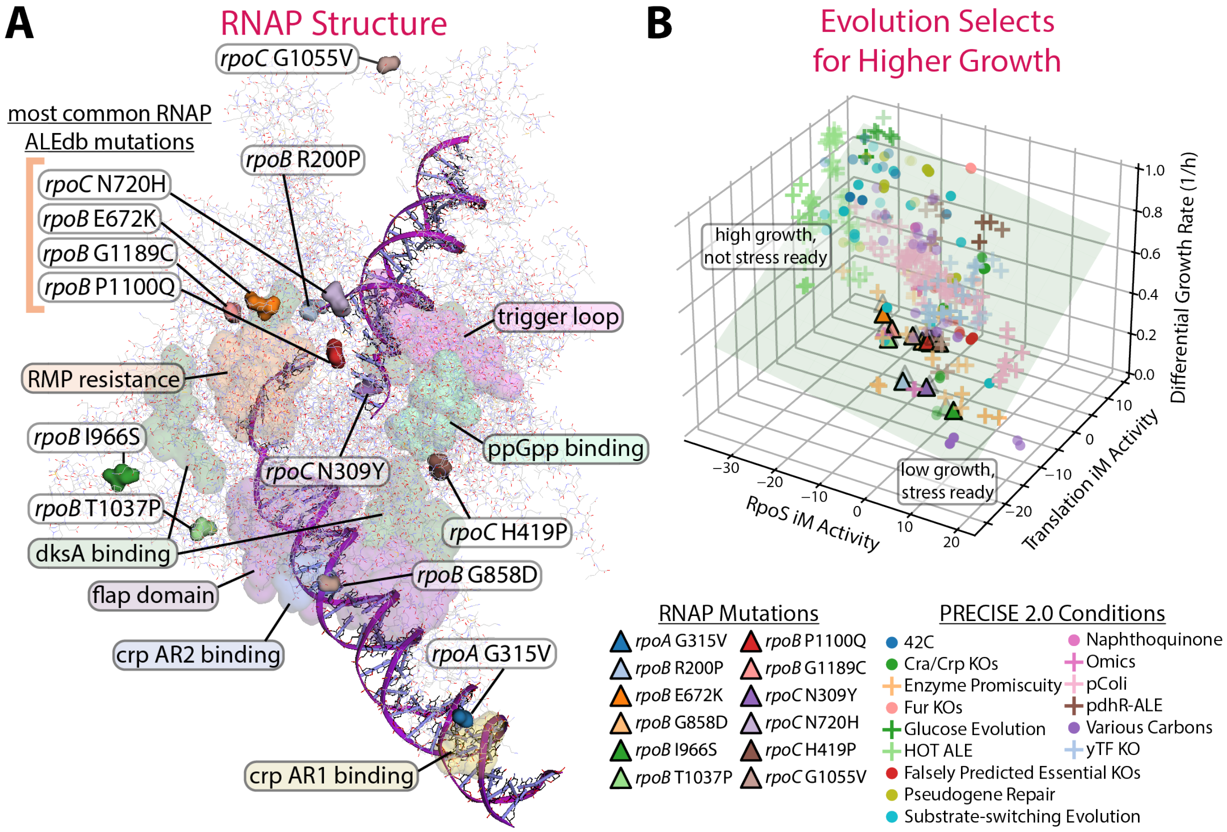

Figure 2.1: **(A)** The structure of RNAP (PDB 6OUL [61]) is visualized using PyRosetta [62], showing the location of mutations used in this study and highlighting some specific RNAP regions of interest [48–56]. The grouped mutations on the upper left are some of the most common mutations found in ALEdb [16] and are further discussed in Figure 2.2. **(B)** Laboratory evolution leads to sequence variants which adjust the composition of the transcriptome leading to faster growth and repressed stress readiness. The f/g tradeoff on the transcriptome is shown (RpoS represents fear, Translation represents greed) along with the mutations' impact on growth rate. All PRECISE 2.0 samples with recorded growth rates are shown. Growth rates are centered on their respective experiments' unevolved control conditions. The green plane is fitted to the data and shows that growth rates increase with lower RpoS and higher Translation iModulon activities.

2.2 Results

2.2.1 Creation of a new dataset of common RNAP mutations

RNAP mutations are frequently fixed in ALEs, with 36% of evolved isolates in ALEdb, a database of mutations acquired during ALE [16], containing at least one RNAP mutation: 6%

have a *rpoA* mutation, 20% have a *rpoB* mutation, and 13% have a *rpoC* mutation. For this study, twelve RNAP mutations were selected and generated for experimental evaluation using three primary criteria: (1) the frequency of occurrence of the mutation in *E. coli* ALE endpoints, (2) their structural location in relation to a known RNAP region of interest, such as effector binding sites, and (3) evidence of phenotypic impact of the mutation. Figure 2.1A shows the location of these twelve mutations on RNAP along with some particular structural regions of interest. Sigma factor binding sites are shown in Supplemental Figure A.2. These mutations were introduced into the genome of the model K-12 MG1655 strain of *E. coli* (see Methods: Creation of RNAP Mutations) to generate single mutation knock-in strains.

RNA-sequencing data was collected under aerobic growth on glucose M9 minimal media for each of these individual mutants (see Methods: RNA-Sequencing). Some of the RNAP mutant strains were additionally tested under specific stress conditions that were similar to the ALE experiment in which they were originally found. All but one of the 12 mutants exhibited a shift toward greed in the f/g tradeoff in the transcriptome (Figure 2.1B). The exception, *rpoB* I966S, arose during an evolution to high temperature growth [63] and may therefore have had a stronger impact on temperature stability than regulation of expression. All but *rpoB* I966S and *rpoC* N309Y, the latter of which arose during butanediol tolerance evolutions, increased the growth rate (Figure 2.2A). *RpoC* N309Y does not increase the growth rate but does shift its transcriptome towards greed in a pattern consistent with the other mutations (Supplemental Figure A.3). It should be noted that *rpoC* N309Y was generated using a different procedure from the other mutations (see Methods: Creation of RNAP Mutations) which could be skewing its results.

2.2.2 Genomic features are patterned unevenly

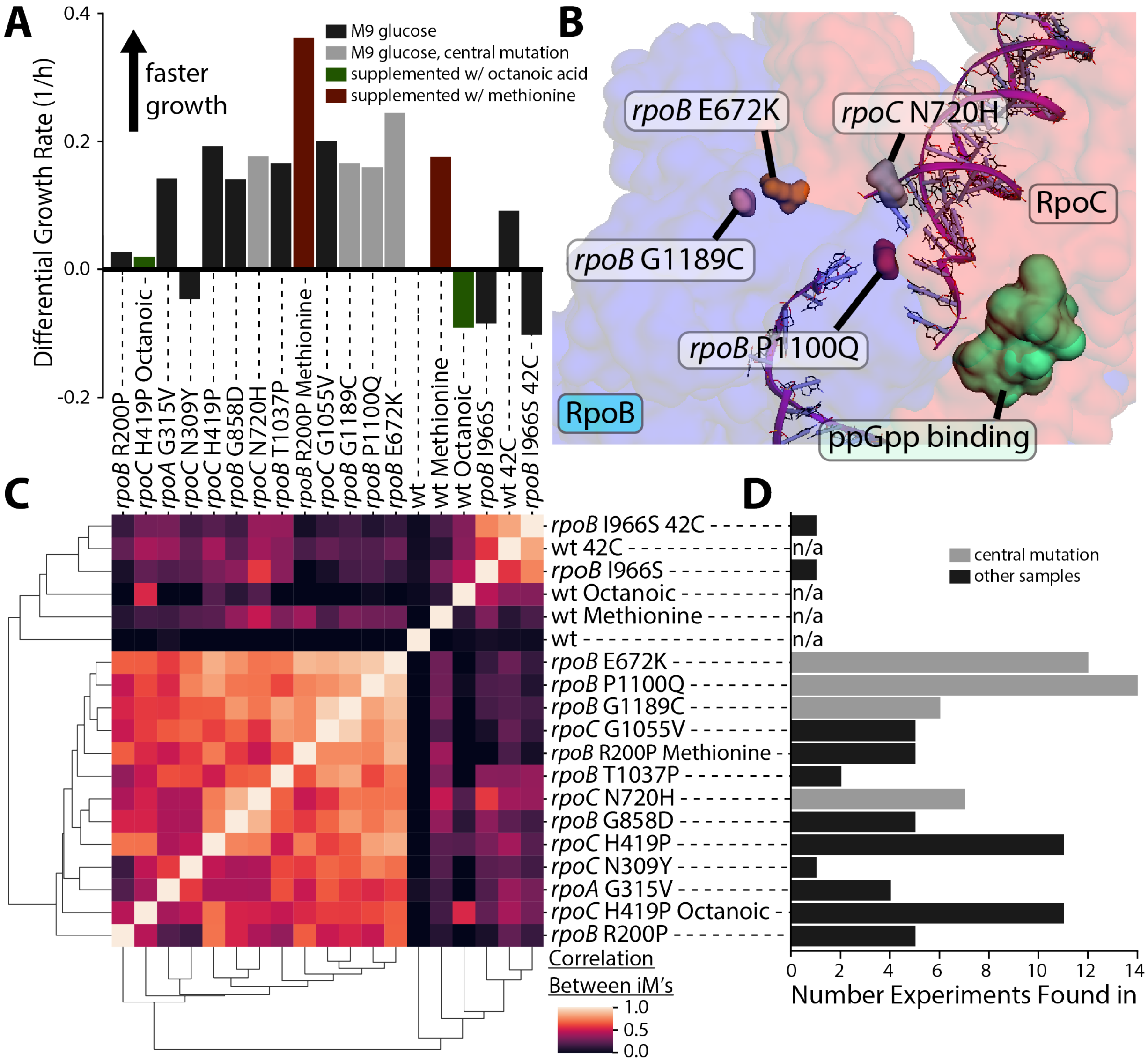

Figure 2.2: **(A)**The growth rates of the mutated strains relative to the wild-type control. **(B)** A subsection of RNAP (PDB 6OUL [61]) showing the location of common mutations with respect to the rpoB-rpoC interface and the ppGpp binding site, visualized using PyRosetta [62]. **(C)** Correlations between the activity levels of all iModulons between RNAP mutants under the same growth condition. The plot shows that all the twelve mutations have a similar impact on transcriptome composition. Mutations in the catalytic core have a near-identical impact on the transcriptome. **(D)** Number of laboratory evolution experiments that RNAP mutations are fixed in (number given is from a total of 743 ALE experiments found in ALEdb [16]). The gray bars in this panel and Panel C are the mutations grouped as "most common RNAP ALEdb mutations" in Figure 2.1 and are visualized on the RNAP structure in Panel B of this figure. [62]

RNAP mutations have been shown to affect RNAP structurally in a variety of ways. Some of the most commonly found and widespread RNAP mutations are *rpoB* E672K, *rpoB* P1100Q, *rpoB* G1189C, and *rpoC* N720H (2.2). The physical mechanism for how these four mutations cause the tradeoff is not fully established, but some key properties are known. Structurally, they are all located near the rpoB-rpoC interface (*rpoB* E672K = 5.46 Å, *rpoB* P1100Q = 5.24 Å, *rpoB* G1189C = 8.97 Å, *rpoC* N720H = 10.09 Å) as visualized in 2.2B. PyRosetta [62] was used to calculate the mean impact of these mutations on the holoenzyme structures and found that all were predicted to destabilize the rpoB-rpoC interface (*rpoB* E672K = -28.40 REU, *rpoB* P1100Q = -23.98 REU, *rpoB* G1189C = -5.26 REU, *rpoC* 1055V = -13.91 REU, mean of all RNAP mutations on ALEdb = -16.83 REU, see Supplemental Figure A.4). This region is nearby to a ppGpp binding site which the mutations are also mostly predicted to destabilize and thus likely modify its regulatory role [51] which is tightly connected to RpoS's own activity [64]. The effect these mutations have on RNAP though are unlikely only limited to the destabilization of said interfaces.

These mutations each may have effects specific to their structural location. *RpoB* E672K for example is located at the base of the bridge helix where it possibly affects DNA-RNAP interactions. *RpoB* P1100Q is near a helix in the beta prime subunit that interacts with ppGpp binding site 1. While some of these mutations are near to ppGpp binding site 1, it should be noted that ppGpp binding site 2 has been reported to have a greater effect on gene expression [51]. DksA, which comprises much of the interface with ppGpp in site 2, has not been mutated in samples found in ALEdb.

Unfortunately a computational structural analysis of how these mutations affect sigma factor binding is not feasible. Sigma factors bind over large portions of RNAP (see Supplemental

Figure A.2) and the specific structural file used has a dominant effect on the resulting destabilization scores. What we can observe though is that the RNAP mutations modify RNAP's interactions with sigma factors nonuniformly. Genes regulated by RpoS [33], the general stress response sigma factor, showed on average a -0.33 change in log2 transcripts per million (tpm) expression when compared to the wild-type. The relatively small change (-0.059 change in log2 tpm) in genes regulated by RpoD, the housekeeping sigma factor, shows that these mutations differentially affect sigma factor functions. This infers that these mutations are preferentially affecting certain sigma factors likely through their binding interfaces.

2.2.3 RNAP mutations lead to upregulation of growth-related genes and downregulation of stress-related genes

The analysis of global changes in the transcriptome is difficult due to the high number of differentially expressed genes in many comparisons. Furthermore, comparing many conditions is challenging if pairwise differential expression of genes (DEG) plots are used [34] (see Supplemental Figure A.5). To overcome these challenges, we used the iModulon workflow [37, 65] to identify independently modulated gene sets (iModulons) and interpret their differential activity between all conditions used. This workflow uses independent component analysis (ICA) of a compendium (X) of RNA-sequencing data, which includes our samples of interest along with a variety of other experiments which help to separate source signals associated with transcriptional regulators [37, 65]. The algorithm generates two output matrices: M (whose columns highlight the genes in each iModulon) and A (whose rows show the iModulon's activity in every sample). Detailed information on each iModulon is available at iModulonDB.org [5] and this study focuses primarily on the "*E. coli* PRECISE 2.0" dataset [65]; an *E. coli* database of RNA-sequencing data obtained under 422 growth conditions. All iModulon activities are measured relative to an unstressed M9 glucose condition and should be interpreted thusly.

Principal component analysis (PCA) of the iModulon activity matrix (A) shows that much of its variance and thus expression variation in general is explained by the RpoS and Translation iModulons' activities. The RpoS iModulon is the largest and the Translation iModulon is the fifth largest contributing factor to the highest variance explaining principal component (PC). GadX and ppGpp iModulons are also highly contributing factors to large variance explaining PC's, adding additional dimensionality to f/g that is further explored in Figure 2.3. The f/g tradeoff is thus a major contributor to variation in the composition of the transcriptome.

The new RNA-sequencing data from the twelve new RNAP mutant strains was analyzed using ICA [37]. The iModulon activity levels in the new samples were compared to those in PRECISE 2.0. This database was used to compute the iModulons structure of the *E. coli* transcriptome [37].

All of the twelve mutations introduced, except for *rpoB* I966S, have a large impact on the activity level of the RpoS iModulon similar to the two previously studied RNAP mutations [30]. The mutations in the catalytic center (those visualized in Figure 2.2B) have the largest impact on RpoS iModulon activity levels (44.1% higher on average than the other RNAP mutants generated for this study as can be seen in Supplemental Figure A.3), but mutations distant from this location can also strongly impact the activity of this iModulon which has not been previously shown. This suggests there is more complexity to the physical mechanism of this transcriptomic effect. Both the frequency of occurrence and the effect of these RNAP mutations found in the catalytic center imply they are commonly fixed during growth rate selection (i.e., maximization of 'greed').

2.2.4 Genome-scale models of proteome allocation quantitatively estimates the growth benefit of maximizing greed functions

While iModulons are an informative approach to reveal the hallmarks of changes in the expression state, they are not directly representative of the composition of the proteome. Creating iModulons from expression data requires the input RNA-sequencing data to be both centered to a control and normalized. This means the activity levels of iModulons for samples are entirely relative to each other and their magnitude range is constrained by the variance of the PRECISE dataset. We thus deployed a genome-scale model to reproduce the f/g tradeoff which allowed us to infer absolute measures of the proteome of cells undergoing said tradeoff. A genome-scale metabolism and expression (ME) model [66] was run to maximize growth while constraining RpoS iModulon-associated reactions to a specified lower bound.

The resulting RpoS and Translation iModulons' proteomic computed mass fractions were highly anticorrelated (-0.9994) (see Supplemental Figure A.6). A unit activity increase in the Translation iModulon has a 650% stronger effect on said iModulons' genes' proteome mass fraction than it does in the RpoS iModulon (see Methods). This implies that the small activity increases of the Translation iModulon seen in the f/g tradeoff and in the RNAP mutations may be having a larger effect than appears on the cell's phenotype. This computational model also indicates that forced expression of the stress readiness genes reduce the expression of the growth promoting genes as experimentally observed.

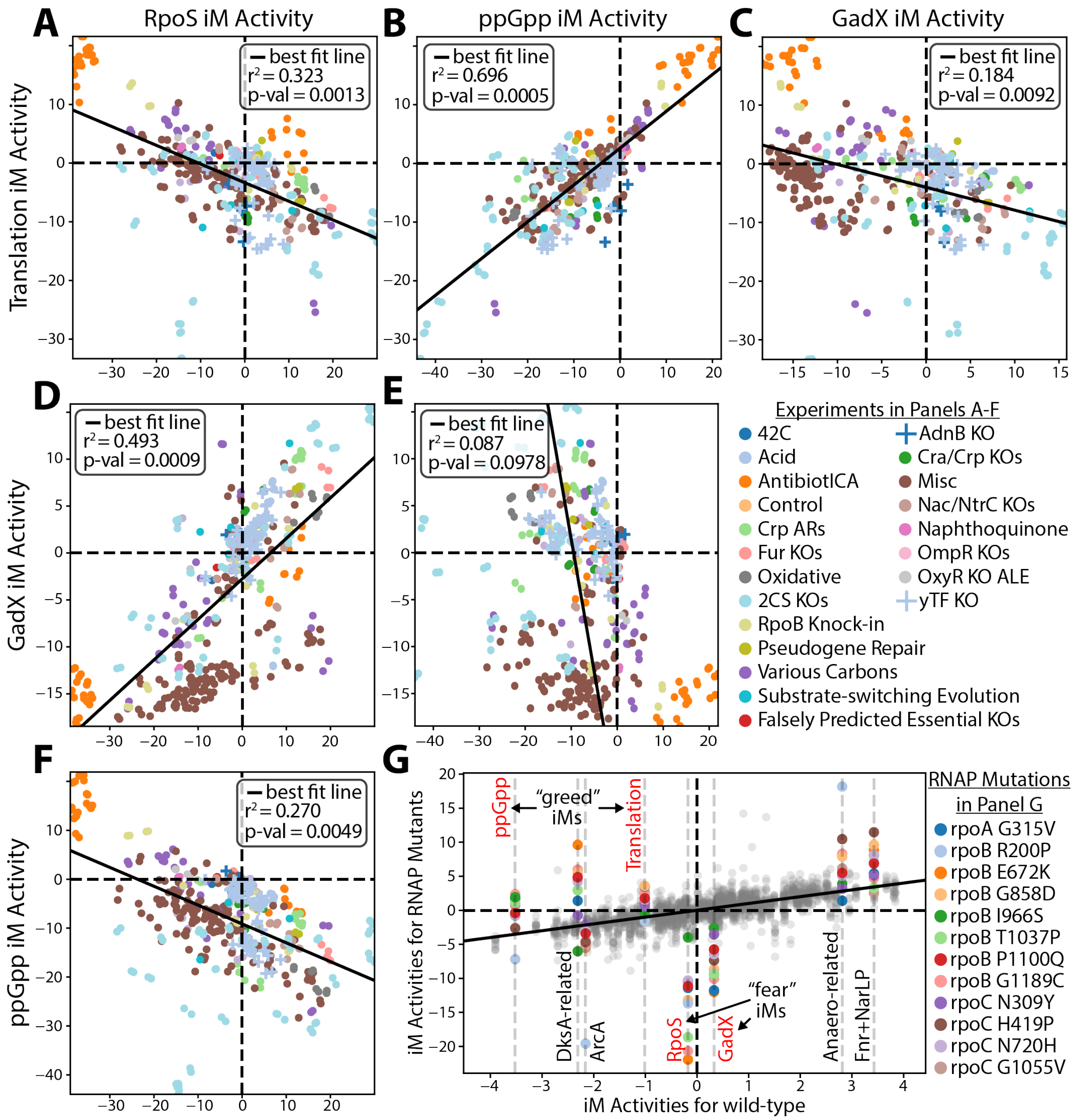

Figure 2.3: **(A-F)** These plots show the relationship in activity levels between the greed (Translation and ppGpp) and fear (RpoS and GadX) iModulons. The p-value is calculated using a t-distribution test of all iModuon-to-iModulon pairwise activity level comparisons. **(G)** The activity levels of various growth- and stress-related iModulons for the RNAP mutants, along with some other iModulons highly affected by said mutations. The gray dots are the activity levels of the other iModulons for all of the mutants. Red labeled iModulons are plotted in panels A-F.

2.2.5 The fear vs. greed tradeoff additionally involves GAD and ppGpp iModulons

The f/g tradeoff was first visualized using the activity levels of the Translation and RpoS iModulons [37]. Since this study was published, the number of transcriptomes for *E. coli* has quadrupled [67]. The analysis of the larger data sets reveals additional dimensionalities to the f/g tradeoff. Several additional iModulon activity levels are correlated with growth rates, including the GadX and ppGpp iModulons. The RNAP mutations are likely affecting these iModulons as ppGpp binds to RNAP while the GAD regulon's expression has been closely tied to RpoS and ppGpp [45, 68]. GadX is highly correlated with RpoS (0.71) and negatively correlated with growth rates (-0.24) while ppGpp is strongly correlated with Translation (0.74) and has a weak positive correlation with growth rates (0.14). Correlation plots for each of these iModulon activity pairings are given in Figure 2.3A-F. All pairings except for GadX and ppGpp iModulons show a clear correlation.

2.2.6 RNAP mutations can be condition-specific adaptations

While the core group of common RNAP mutations downregulate stress-related iModulons and upregulate growth-related iModulons (Figure 2.3G), other RNAP mutations have more specific effects that are adaptations to the environments from which they were found. Supplemental Figure A.7 shows two of these such mutations (*rpoB* R200P and *rpoA* G315V) from our set of twelve mutations.

The *rpoB* R200P mutation reflects a specific selection condition. It is found commonly in replicate methionine tolerance evolutions [59] and it has two effects on the transcriptome: (1) during growth on methionine it activates the Translation iModulon and downregulates the

RpoS iModulon to increase the growth rate compared to wild-type; and (2) during growth on M9 glucose it activates anaerobic response genes found in the Fnr-1, Fnr-2, and Anaero-related iModulons. These responses are likely used because methionine contains sulfur and is thus a common target of reactive oxygen species (ROS) in *E. coli* [69].

The *rpoA* G315V mutation affects the activities of Crp-1 and Crp-2 iModulons with the strongest impact on the maltose operons. This mutation was found in a *pgi* synthetic gene replacement ALE [58] in nearly all strains that failed to integrate the exogenous *pgi* replacements. Presumably the loss of *pgi* required large changes to sugar import systems, thus necessitating this *rpoA* mutation to help downregulate maltose importers [70]. The mutation's effect on the Crp-1 iModulon is similar to one reported in a study that deactivated regions of crp [71] (see Supplemental Figure A.7E). Thus it is likely the mechanism of action for this *rpoA* mutation is to modify the rpoA-crp binding interface.

Thus, there are RNAP mutations outside the core of the enzyme that confer condition-specific effects on the transcriptome (see A for more cases). This observation leads to a wider examination of the effects of RNAP mutations that are selected for under specific conditions.

2.2.7 The genetic basis for the fear vs. greed tradeoff during ALE is condition-dependent

The primary fear and greed iModulons are correlated for both unevolved samples (-0.57 correlation) and evolved samples (-0.39 correlation, see Figure 2.4A), although evolved samples strongly favor greed. These correlations hold true across samples with and without RNAP mutations, but RNAP mutations nearly universally favor a movement towards greed. Different stressors lead to specific transcriptional adjustments along the f/g tradeoff to best favor growth

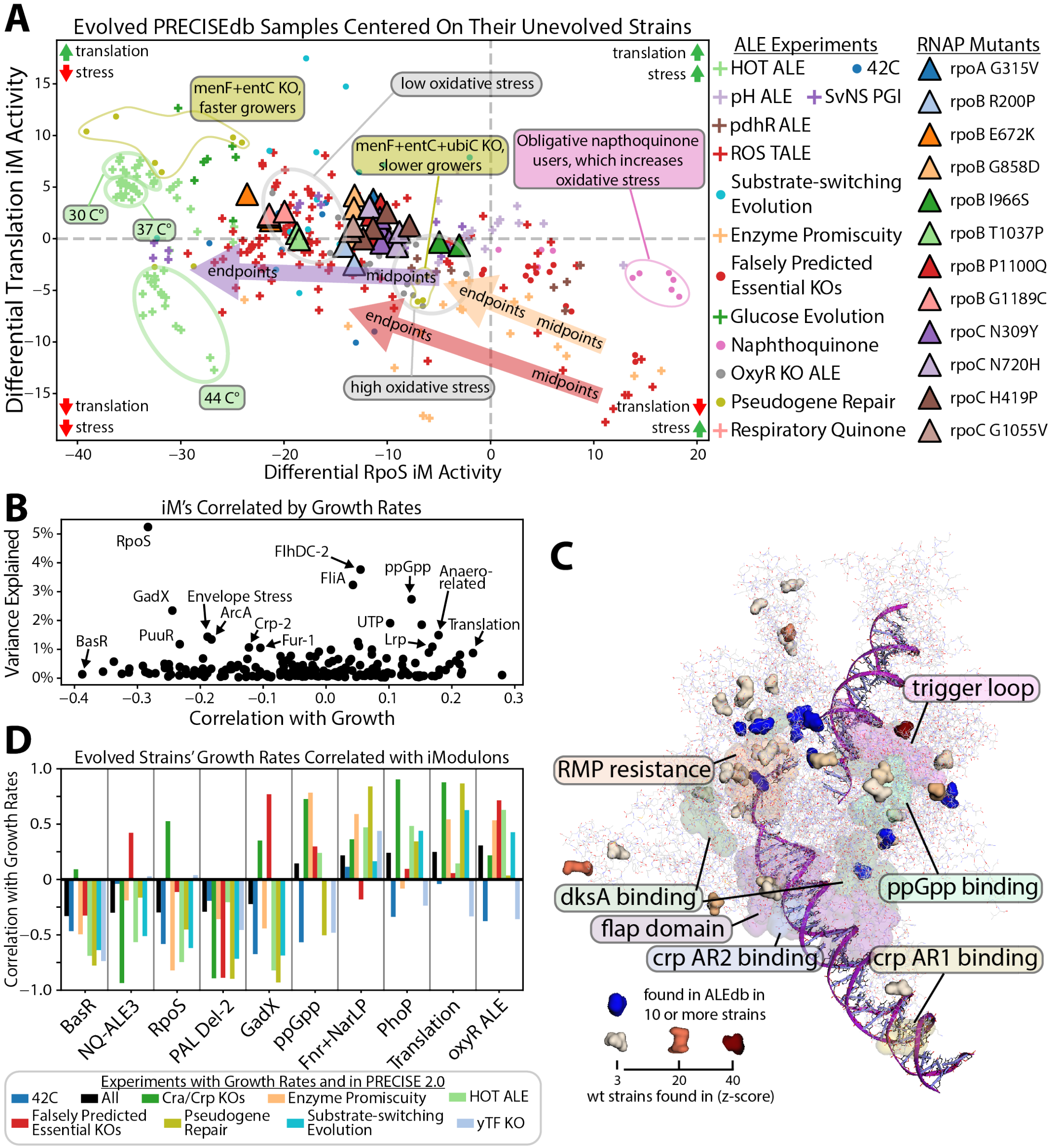

Figure 2.4: **(A)** The fear vs. greed iModulon activities of the evolved samples of PRECISE 2.0 centered on their respective unevolved wild-type strains' iModulon activities. **(B)** The iModulons of PRECISE 2.0 correlated with the available growth data along with how much of the total transcriptome's variance they explain. **(C)** The most common mutations found in ALEdb and the natural variants of RNAP (PDB 6OUL [61]) visualized using PyRosetta [62]. **(D)** The correlation values between growth rates and iModulons for all evolution experiments with growth rates reported.

as is annotated in Figure 2.4A.

In most laboratory evolutions with high stress conditions, evolution downregulates the RpoS iModulon over time. The cells initially use the RpoS iModulon to respond to nearly any stress, but eventually tune the stress response to the specific environment. In a reaction oxygen species experiment (labeled ROS TALE) [72], initially the RpoS iModulon was highly active but as the cells evolved on paraquat most of the iModulon was downregulated while the expression of oxidative response genes in the iModulon were left largely unmodified (see Supplemental Figure A.8). This transcriptional regulatory network adjustment, which was driven by convergent mutations to icd, aceE, sucA, oppA, and emrE among other genes, enabled the cells to grow faster in a ROS stress environment.

2.2.8 Growth rates are well correlated with the fear vs. greed tradeoff

Standardizing growth rates across experiments is a difficult task, as unintentional differences in laboratory procedures, data processing, or simple measurement bias can drastically skew the results while intentional differences in experimental conditions make a direct comparison difficult. Growth rate data is also often not reported, as just 43% of PRECISE 2.0 samples have associated growth rates. The growth data present, however, supports the fear vs. greed tradeoff. Translation is the second most positively correlated iModulon with growth while the RpoS iModulon is the tenth most negatively correlated with growth (Figure 2.4B). It is important to note that the iModulons with stronger correlations to growth than Translation and RpoS explain little of the transcriptome's variance. Figure 2.4D shows that these correlations hold true across a variety of evolution experiments.

OxyR ALE, for example, is the most correlated positively iModulon with growth rates yet explains only 0.1% of the transcritome's variance and its activity is nearly entirely limited to the ALE study for which it is named. Anaero-related, in addition to Translation and ppGpp, has a positive correlation with growth and a large explained variance of PRECISE's expression data. While it also is upregulated by the RNAP mutations, compared to the other greed iModulons it contains many genes of unknown function and has no clear regulator. The f/g tradeoff is defined not by all growth and stress related genes but rather key well-defined stress and growth iModulons whose activities anti-correlate with each other over a large range of conditions. However, future versions of PRECISE will likely enable the inclusion of the Anaero-related iModulon among others into the f/g tradeoff.

This ceaseless pull towards greed and away from stress readiness, however, is largely limited to laboratory conditions. The lack of overlap between the natural variants [73] and the ALEdb mutations seen in Figure 2.4C implies that there are highly divergent evolutionary pressures on wild-type strains and their ALE counterparts.

2.2.9 The fear vs. greed tradeoff is found across the phylogenetic tree

Finally, we searched the phylogenetic tree for other organisms exhibiting the f/g tradeoff (the phylogenetic tree highlighting said species can be seen in Supplemental Figure A.9). First we analyzed data from a multi-strain *E. coli* ALE study [75]. This analysis shows that the tradeoff was found in all the *E. coli* strains of the study (Figure 2.5A). Second, we examined iModulonDB [5] for the presence of the f/g in other species (Figure 2.5D-K). The tradeoff was clearly found in seven out of the 12 bacterial strains surveyed (see Methods: Cross-species iModulon Comparisons). Although the gene composition of the fear iModulons varies between species (likely a consequence of differing stresses in their natural environments), all of the primary greed

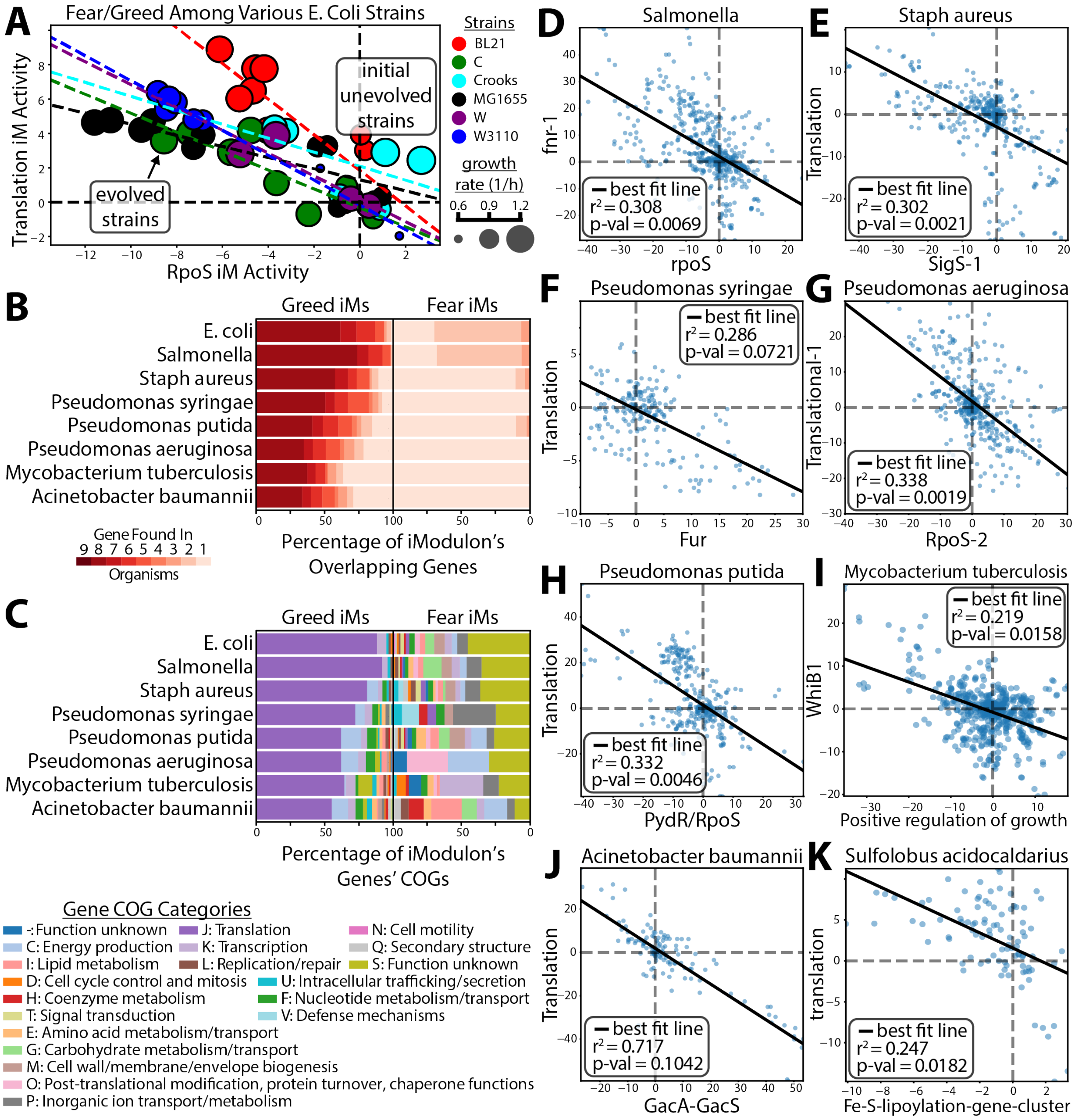

Figure 2.5: **(A)** The f/g tradeoff appears in ALEs across multiple *E. coli* strains [74]. **(B)** Percentage of genes found in common among translation and stress iModulons in different species. **(C)** The COG category of the genes of the greed and fear iModulons. **(D-K)** The f/g tradeoff among a variety of species found in iModulonDB [5]. The p-value is calculated using a t-distribution test of all iModuon-to-iModulon pairwise activity level comparisons. The names of the iModulons are pulled from their respective data sets. Mycobacterium tuberculosis' "Positive regulation of growth" iModulon mostly consists of stress-related antitoxin genes.

iModulons consist of a highly similar set of ribosomal subunits and translational associated functions (Figure 2.5B-C). The five species in which the tradeoff was not found all contain a greed iModulon that consists primarily of ribosomal subunits, but said species contain no one clear stress iModulon that correlates with it. The presence of the greed-related genes of the f/g tradeoff across such a wide range of species implies that they may be a global property of bacterial transcriptomes.

2.3 Discussion

We detail a general tradeoff in the bacterial transcriptome between growth rate and stress readiness. A major genetic component of this tradeoff lies in RNAP mutations, which affect the structure of RNAP and consequently the composition of the transcriptome. In RNAP mutants that arise from ALE studies, the modified transcriptome composition favors transcription of growth-related functions over stress-related functions. The tradeoff between fear and greed related functions was found across a wide range of wild-type bacterial strains. Similar transcriptional tradeoffs have been seen before in persistence [76], nutritional competence [77], and protein cost in metabolic pathways [78]. Interestingly, the fear vs. greed tradeoff has been described in many areas of science; such as economics [79], game theory [80], and psychology [81]. It has been elucidated here for microbiology through a multi-scale analysis.

A previous study compared two RNAP mutations [30], *rpoB* E672K and *rpoB* E546V, and found that they destabilize the rpoB-rpoC interface [82]. Another study using in vitro assays linked an *rpoC* deletion from 3,611 to 3,619 bp (near to the rpoB-rpoC interface) to destabilizing the open complex of RNAP which led to decreased transcriptional pausing on the promoter, reduced RNAP's open complex half-life, and increased elongation rates [24]. For our centrally located mutations, our evidence best supports this model of a destabilized rpoB-rpoC interface leading to a destabilized open complex thus causing transcriptional changes. However, we have no clear mechanistic explanation as to why mutations distant from this central region, such as *rpoC* G1055V, have similar impacts to the transcriptome. The impact of RNAP mutations have also been shown to be similar to strains with reduced number of ribosomal operons, suggesting that these mutations are possibly modifying ribosomal availability and/or distribution [4]. Other RNAP mutations were found to eliminate the destabilizing effect of ppGpp binding to RNAP, thus reducing the inhibition of transcription by ppGpp [83].

A recent study analyzing 45,000 ALE mutations and comparing them to wild-type variant alleles suggests that under laboratory evolution the wild-type alleles are under negative selection pressure, while ALE mutations are under positive selection pressure [73]. This suggests that ALE mutations represent extreme mutations extenuating a preferred trait, thus amplifying the basis for the f/g tradeoff as opposed to nature in which a sole focus towards faster growth would leave cells unable to adapt to highly variable conditions.

The current study expanded upon current knowledge [24, 30] by analyzing the impact of twelve RNAP mutations to detail RNAP's role as a global master regulator of the f/g tradeoff. All twelve of these mutations, however, are from evolution experiments and their common adjustments towards greed are reflective of that. The detailed molecular/structural mechanisms that underlie the tradeoff are not fully understood, but appear to involve the rpoB-rpoC interface [30] and other important structural regions of RNAP, altered kinetic and regulatory properties [24], and changes in the sigma factor use of RNAP.

The effects that RNAP mutations have on the transcriptome composition, however, are clear. The transcriptomic re-allocation involves a consistent set of iModulons with known functions. As additional versions of PRECISE are created using more data, it is likely additional iModulons could be included in this tradeoff. The relationship between the proteome and transcriptome functions enable genome-scale computational biology assessment of the phenotypic consequences of the reallocation [84]. Thus, a detailed understanding of the effects of the f/g tradeoff at the systems level has emerged. As the tradeoff involves resource allocations for improved fitness, it is important to contextualize particular RNAP mutations fixed in laboratory evolution studies and seek to identify adaptive mutations that are condition specific.

Finally, the phylogenetic distribution of the greed-related genes of the f/g tradeoff is broad, suggesting that this tradeoff may emerge as a universal feature of the bacterial transcriptome that can be captured by iModulons. It is not known, however, if RNAP mutations would have a similar impact to the tradeoff in these species. The tradeoff has been also found in a minimal synthetic organism, further supporting its potential ancient origin [85]. RNAP and the f/g tradeoff have been shown to play a highly important role in balancing growth and stress adaptations.

2.4 Methods

Strain Information

E. coli K-12 MG1655 was used as the wild-type and as the source strain for all mutations created for this study.

Creation of RNAP Mutations

RpoC N309Y was created using pORTMAGE [86], the protocol for which is included in Supplemental File 1 of the publication. Initially pORTMAGE was intended to be used to generate all strains, but only *rpoC* N309Y could successfully be generated and thus the rest were created using the CRISPR-based protocol outlined in Zhao et al. [87]. Mutations were verified using reverse PCR. Primer sequences used in the generation of mutants are included in Supplemental File 2 of the publication.

Growth Rate Calculations and Comparisons

Reproductive growth rates were calculated under the same conditions for all RNAP mutated strains. A 24-well magnetic heat lock set to 37° C was used for continual cultures. 16mL culture with OD600 = 0.05 using M9 minimal media supplemented with 4 g/L glucose was prepared in a plastic tube; and time points taken in replicate for growth rate calculation approximately every 30 minutes. For comparing growth rates across experiments, all growth rates were analyzed as differential values relative to their respective experiment's control condition. After being centered on their respective control conditions, the differential growth rates were normalized for each experiment. The growth rate values for the PRECISE 2.0 samples are available at iModulonDB (https://imodulondb.org/organisms/e_coli/precise2/data_files/sample_table.csv).

RNA-sequencing

All samples were prepared and collected in biological duplicates. 3 ml of culture isolated at an OD600 of 0.5 was added to 6 ml of Qiagen RNA-protect Bacteria Reagent after sample collection. This solution was then vortexed for 5 seconds, incubated at room temperature for 5 minutes, and then centrifuged. The supernatant was then removed and the cell pellet was stored at -80° C. The Zymo Research Quick-RNA MicroPrep Kit was used to extract RNA from the cell pellets per vendor protocol. On-columns DNase treatment was performed for 30 minutes at room temperature. Anti-rRNA DNA Oligo mix and Hybridase Thermostable RNase H [88] was used to remove ribosomal RNA. Sequencing libraries were created using a Kapa Biosystems RNA HyperPrep per vendor protocol. RNA-sequencing reads were processed using https://github.com/avsastry/modulome-workflow. Data is available at NCBI GEO GSE227624.

iModulon Computation

RNA-sequencing data was used to create iModulon activity levels of the mutated strains using PyModulon [37] which is available at https://github.com/SBRG/pymodulon. Activities of iModulons were compared to samples from PRECISE 2.0 [65] which is easily accessible using iModulonDB [5].

Mutation Analysis

ALEdb [16] was used for selecting the mutations for this study. Any *E. coli* strains on ALEdb were considered as potential sources for mutations. Mutations from the same sample but where one is from an isolate and one is from the population were considered to be just one instance of said mutation.

Structural Analysis

Structural analysis was performed using PyRosetta [62] using its default score function. The pdb files were downloaded from RCSB [89]. REU stands for Rosetta Energy Unit, which is PyRosetta's unit for energy. The files used were selected primarily based on a review of bacterial RNAP [90].

The structural calculations for the rpoB-rpoC binding interface were performed by calculating the binding energy between the chains coded by *rpoB* and *rpoC* using the holoenzyme pdb structures. For each mutation, said mutation was introduced, the protein was repacked, and the binding energy between the two chains was recalculated and compared to the baseline. The structural calculations for ppGpp binding analyses were carried out similar to the rpoB-rpoC binding simulations, but by instead calculating the binding energy between ppGpp and the rest of the protein.

Structural analysis was carried out for each of the twelve mutations created specifically in this study (see Supplemental Fig. A.4). Calculations were also carried out for an alanine scan of RNAP and all ALEdb RNAP mutations to serve as various controls (see Supplemental Fig. A.10).

Data Processing

iModulons are calculated using expression data centered on a control. For this study the control was a wild-type M9 glucose growth sample on which all other samples were centered. All biological replicates of expression data had over 99% correlation to each other and were averaged together. In addition to PyRosetta [62] and PyModulon [37], numpy [91], pandas [92], and scipy [93] were used to generate figures and perform analysis.

Metabolic Model and Proteomic Calculations

The FoldME [66] model was used for the metabolic modeling calculations. Supplemental Figure A.6 was generated by iteratively increasing the lower bounds for the genes of the RpoS iModulon and recording the proteomic mass fraction of the Translation and RpoS iModulons' genes until the model no longer ran. Proteome mass fraction to iModulon genes is the sum of the measured proteomic mass fractions of each enriched gene in an iModulon. This value is calculated for every sample and plotted against its corresponding PRECISE iModulon Activity. The proteomic calculations performed for this paper are well described in Patel et al [84].

Cross-species iModulon Comparisons

To compare iModulons across different species, first genes from the various strains were matched to each other using Orthofinder on its default settings [94]. The FASTA files for each organism were pulled from their respective NCBI genome pages and fed into the algorithm. The many-to-many Orthofinder results were used to generate the gene mapping for later steps. In the case that an organism had multiple genes mapped to one orthogroup, the multiple genes' weightings were averaged when mapped to the orthogroup. The many-to-many results were used based on the rarity of one-to-one orthologs and at the suggestion of Orthofinder's GitHub page.

Species' iModulons were mapped to both *E. coli*'s Translation and RpoS iModulons based on Fisher's exact test p-values generated on orthogroup presence/absence in iModulons. Said presence/absence calls were generated using k-means clustering of the orthogroup activity levels within iModulons with the number of clusters set to 2 and taking the smaller cluster as the orthogroups present in an iModulon. The iModulon from each species with the lowest p-values were selected as the best matching iModulon. In the case where no iModulon matched the *E. coli* RpoS iModulon (namely for Mycobacterium tuberculosis and Acinetobacter baumannii), the iModulon most negatively correlated with their Translation iModulon was chosen. iModulon names were pulled from the individual species' iModulons.

Acknowledgments

This work was supported by The Novo Nordisk Foundation (NNF) Center for Biosustainability (CfB) at the Technical University of Denmark (NNF20CC0035580), National Institutes of Health Grant R01 GM057089, and the Y.C. Fung Endowed Chair in Bioengineering at the University of California San Diego. We would like to thank Marc Abrams (Systems Biology Research Group, University of California San Diego) for assistance with paper editing, José Utrilla (Center for Genomic Sciences, UNAM) for help with historical context, and Hyungyu Lim (Inha University) for generating the stationary phase data used in Supplemental Figure A.11.

Chapter 2 in part is a reprint of material published in:

- **C Dalldorf**, K Rychel, R Szubin, Y Hefner, A Patel, DC Zielinski, BO Palsson. 2024. "The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions" *mSystems*, 10.1128/msystems.00305-24. The dissertation author was the primary author.

Chapter 3

Diversity of transcription regulatory adaptation in *E. coli*

The Transcriptional Regulatory Network (TRN) in bacteria is thought to rapidly evolve in response to selection pressures, modulating transcription factor (TF) activities and interactions. In order to probe the limits and mechanisms surrounding the short-term adaptability of the TRN, we generated, evolved, and characterized knockout (KO) strains in *E. coli* for 11 TFs selected based on measured growth impact on glucose minimal media. All but one knockout strain (Δ*lrp*) were able to recover growth and did so requiring few convergent mutations. We found that the TF knockout adaptations could be divided into four categories: 1) Strains (Δ*argR*, Δ*basR*, Δ*lon*, Δ*zntR*, Δ*zur*) that recovered growth without any TF-specific adaptations, likely due to minimal activity of the TF on the growth condition, 2) Strains (Δ*cytR*, Δ*mlrA*, Δ*ybaO*) that recovered growth without TF-specific mutations but with differential expression of regulators with overlapping regulons to the KO'ed TF, 3) Strains (Δ*crp*, Δ*fur*) that recovered growth using convergent mutations within their regulatory networks, including regulated promoters and connected regulators, and 4) Strains (Δlrp) that were unable to fully recover growth, seemingly due to the broad connectivity of the TF within the TRN. Analyzing growth capabilities in evolved and unevolved strains indicated that growth adaptation can restore fitness to diverse substrates often despite a lack of TF-specific mutations. This work reveals the breadth of TRN adaptive mechanisms and suggests these mechanisms can be anticipated based on the network and functional context of the perturbed TFs.

3.1 Background

The bacterial Transcriptional Regulatory Network (TRN) adapts to environmental changes [95] primarily through the action of transcription factors (TFs) [96]. Research into TFs has revealed insights into their regulatory targets [97], mechanisms of action [98], and roles in environmental responses [95]. Genetic mutation enables flexibility of TF activities and interactions both on long [95] and short (¡1000 generations) timescales [99]. TF-related mutations have been observed in laboratory evolution experiments [42, 100, 101], indicating that microbes can meet short-term environmental challenges through genetic modulation of the TRN. An investigation of how the TRN adapts in the short term to recover growth following strong perturbations would provide insights into both the mechanistic basis and limits of the plasticity of the TRN.

Knockout adaptive laboratory evolution (KO-ALE) has previously been used to study evolutionary adjustments to genetic perturbations in *E. coli.* Metabolic gene KO-ALEs revealed multiple optimal phenotypes exist to alleviate bottlenecks created by the KO, all of which substantially recover growth [102, 103]. Certain TF KO-ALE experiments have been previously carried out, including a *crp* KO-ALE that found convergent mutations to *ptsG*, an important gene in the glucose phosphotransferase system [100], and a *pdhR* KO-ALE that resulted in targeted mutations to the Shine-Dalgarno sequences of genes in *pdhR*'s regulon [101]. Mutations to regulators themselves have also been found when the TRN is perturbed in other ways, such as an enrichment of *crp* mutations following the KO of adenylate cyclase that produces the *crp* effector cAMP [104] and convergent mutations to *oxyR* following evolution with oxidative stress [42]. In addition to these specific KO-ALE studies, another grew the Keio collection on M9 minimal media supplemented with glucose and found growth defects in KOs of TFs with no known function under this condition [105].

To obtain a more comprehensive understanding of short-term TRN plasticity, we performed the largest to-date study of the adaptive response to the removal of transcription factors in *E. coli*. We carried out KO-ALEs for 11 TFs, each with six independent lineages. The resulting midpoint and endpoint strains were sequenced, expression profiled, and characterized for substrate readiness using phenotyping plates. Independent component analysis (ICA) was used to analyze the gene expression response of the resulting strains, empowering the analysis through comparison to a broad range of experimental conditions in the PRECISE 1K *E. coli* gene expression database [37, 106] (Figure 1A). The results of these experiments suggest a mutation landscape and network structure that is capable of rapidly responding to large perturbations and reveal the breadth of mechanisms underlying the adaptability of the *E. coli* TRN.

3.2 Results

3.2.1 Selection of Transcription Factors for KO and Laboratory Evolution

In selecting TFs for KO-ALE, we sought deletions that would have a functional impact and therefore elicit targeted adaptations. We also prioritized TFs that have well characterized regulons to enable the estimation of condition-specific TF activity. To estimate TF activity, we utilized data-derived transcription modules called iModulons, which are machine learning-computed transcription regulatory modules that have been extensively analyzed on thousands of experimental conditions [106]. TFs were therefore selected for KO-ALE based on three primary criteria (visualized in Figure 3.1A): the regulon size based on binding sites annotated in regulonDB [107], the growth rate impact of the KO on glucose M9 minimal media [105], and how well the TF-associated iModulon explains the variance in expression of genes in the TF regulon [108]. These criteria resulted in 13 targets, 11 of which showed at least a 20% initial growth defect and were subsequently selected to move forward with ALE. Each of these 11 KO strains were evolved in 6 separate lineages with midpoint and endpoint evolution isolates taken from each lineage. All isolates were sequenced to identify genetic mutations, the results of which informed the selection of a subset of the isolates to be expression profiled.

3.2.2 Regulator activity on the evolved growth condition is a primary determinant of ALE dynamics

To anticipate the degree of impact of each TF KO, we classified TFs based on activity on glucose minimal media (Figure 3.1B). We applied a mathematical transformation to adjust all iM activities to have 0 as their minimum value (see Methods: Basal iModulon Transformation). Based on the transformed activity levels of our wildtype control samples and the genetic

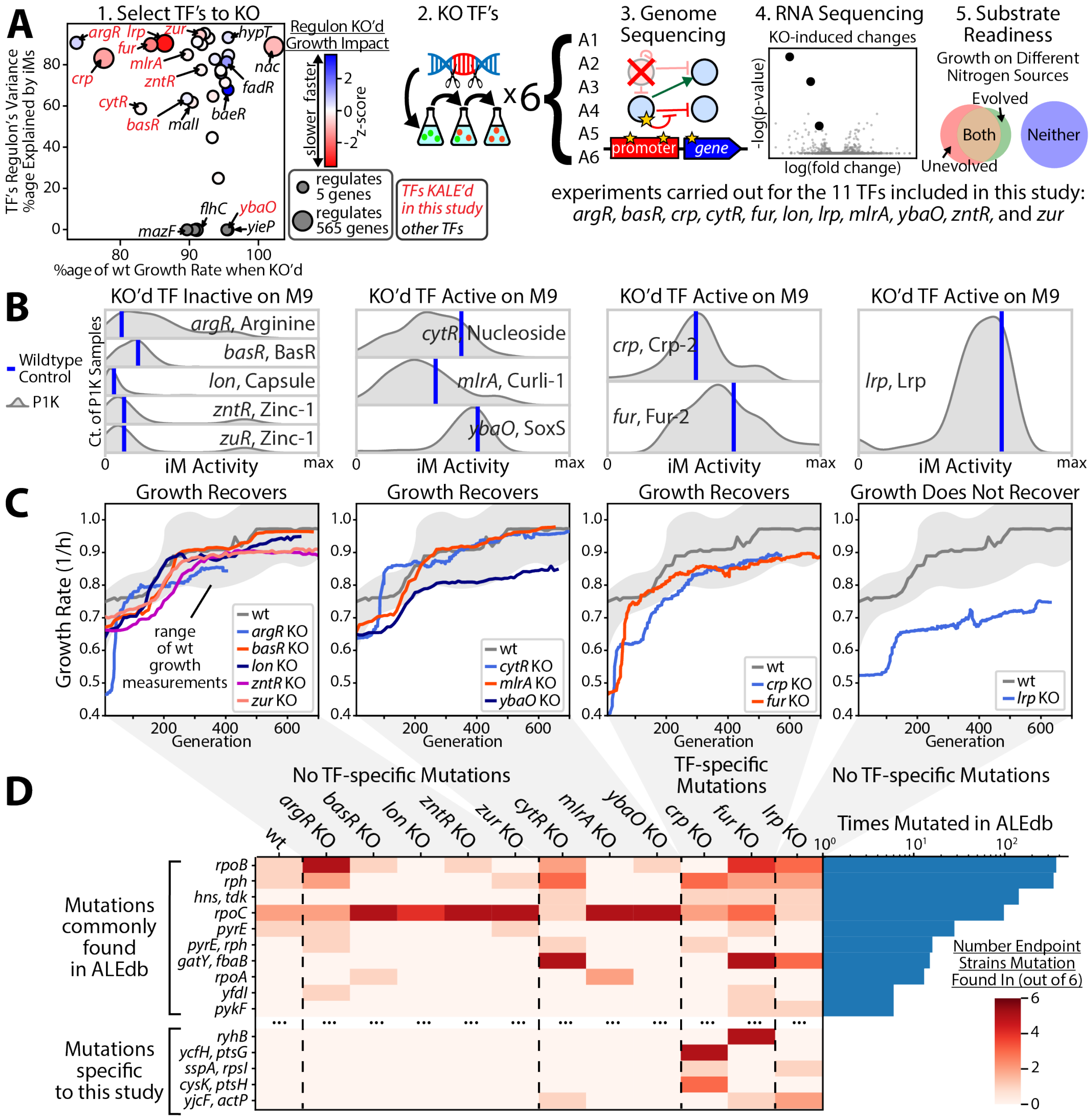

Figure 3.1: **(A)** Eleven TFs were selected for KO-ALE. With the exception of *nac*, only TFs with a negative impact to growth when removed are shown [105]. The coloring for the leftmost figure is set for each TF based on the combined impact to growth rate of all genes repressed by the TF subtracted by the combined impact of all genes promoted by the TF. **(B)** The activity level of the iModulons regulated by the TF KO's. The gene and an iModulon it regulates are named in each subplot (gene, iModulon). These iModulon activities are corrected for basal activity, where zero represents no iModulon activity and all activity values are positive. **(C)** Growth profiles for each of the KO's. **(D)** A mutation table showing the genes mutated across the strains in this study. The right histogram shows how many instances of each mutation can be found in ALEdb (including this study).

adaptations seen in the evolved strains, we classify the TF KO strains into four primary categories that can be seen in Figure 3.1BCD: 1 - TF is inactive on M9 and growth recovers through non-TF specific mutations (*argR*, *basR*, *lon*, *zntR*, and *zur*), 2 - TF is active on M9 and growth recovers through non-TF specific mutations (*cytR*, *mlrA*, and *ybaO*), 3 - TF is active on M9 and growth recovers through TF specific mutations (*crp* and *fur*), and 4 - TF is active on M9 and there is poor growth recovery with non-TF specific mutations (*lrp*). We discuss each of these categories below. Unless otherwise specified, all other figures and analyses outside of this classification step use the standard PRECISE1K (P1K) [106] version of iModulons publicly available at iModulonDB.org.

3.2.3 Almost all TF knockout strains can recover growth with only few mutations

The growth rates seen in Figure 3.1C show that all of the KOs, except for *lrp*, nearly fully recover their growth rates. The *ybaO* KO strains were considered to have recovered growth as their growth data overlaps with the wildtype growth measurements which *lrp*'s growth data does not (see Supplemental Figure B.1). All of the TF KO-ALE strains except for crp and fur did not exhibit any TF-specific causal mutations throughout their evolution (Figure 3.1D). For these strains, mutations were largely limited to non-convergent RNA polymerase (RNAP) mutations and expression differences primarily consisted of the downregulation of stress-related genes contained in the RpoS iModulon and the upregulation of ribosomal subunits contained in the Translation iModulon (Supplemental Figure B.2). These adaptations are common across laboratory evolution experiments [109].

3.2.4 Adaptation to removal of active TFs with small regulons occurs through compensatory mechanisms

For the majority of TF KOs, the expression changes in their evolved strains were not genes regulated by the KO and the TF's regulon was not enriched for mutations (see Supplemental Figure B.3). *BasR*, for example, is the primary regulator of its eponymous iModulon which is largely unchanged during both the knockout and subsequent evolution (Figure 3.2A). The BasR iModulon contains few other regulators (Figure 3.2B), but as *basR* is not normally active on M9 the *basR* KO has little effect on the iModulon's activity. Of minor interest is that *basS*, the sensor kinase for *basR*, is extremely highly expressed in the *basR* KO samples although we have no clear explanation as for what effect this would have on the transcriptome. The lack of clear KO-specific expression differences or convergent mutations in these inactive TF KO-ALE samples infers they are using non-genetic mechanisms to recover from their initial growth defect.

The Curli-1 iModulon, which is regulated by *mlrA*, is active on M9 but the *mlrA* KO and subsequent evolution has little effect on its expression (Figure 3.2C). This is possibly due to the fact that Curli-1 contains many other regulators, some of whom are differentially expressed in the *mlrA* KO samples (Figure 3.2D). It appears that other nearby regulators can help adapt to the loss of a TF in order to maintain normal iModulon activity levels.

The iModulon pipeline [110] was run using the P1K dataset with samples from this study added, which resulted in new iModulons (Figure 3.2E). There is a KO-specific iModulon for all but one of the TF KO-ALEs with expression profiles (185 for the *basR* KO iModulon, 93 - *crp*, 220 - *fur*, and 114 - *lrp*) which is common in KO studies. The genes in KO-specific iModulons often have fewer regulators than genes in most iModulons (Figure 3.2F), showing that the lack of other nearby regulators leaves these genes largely unregulated which ICA captures as KO-specific

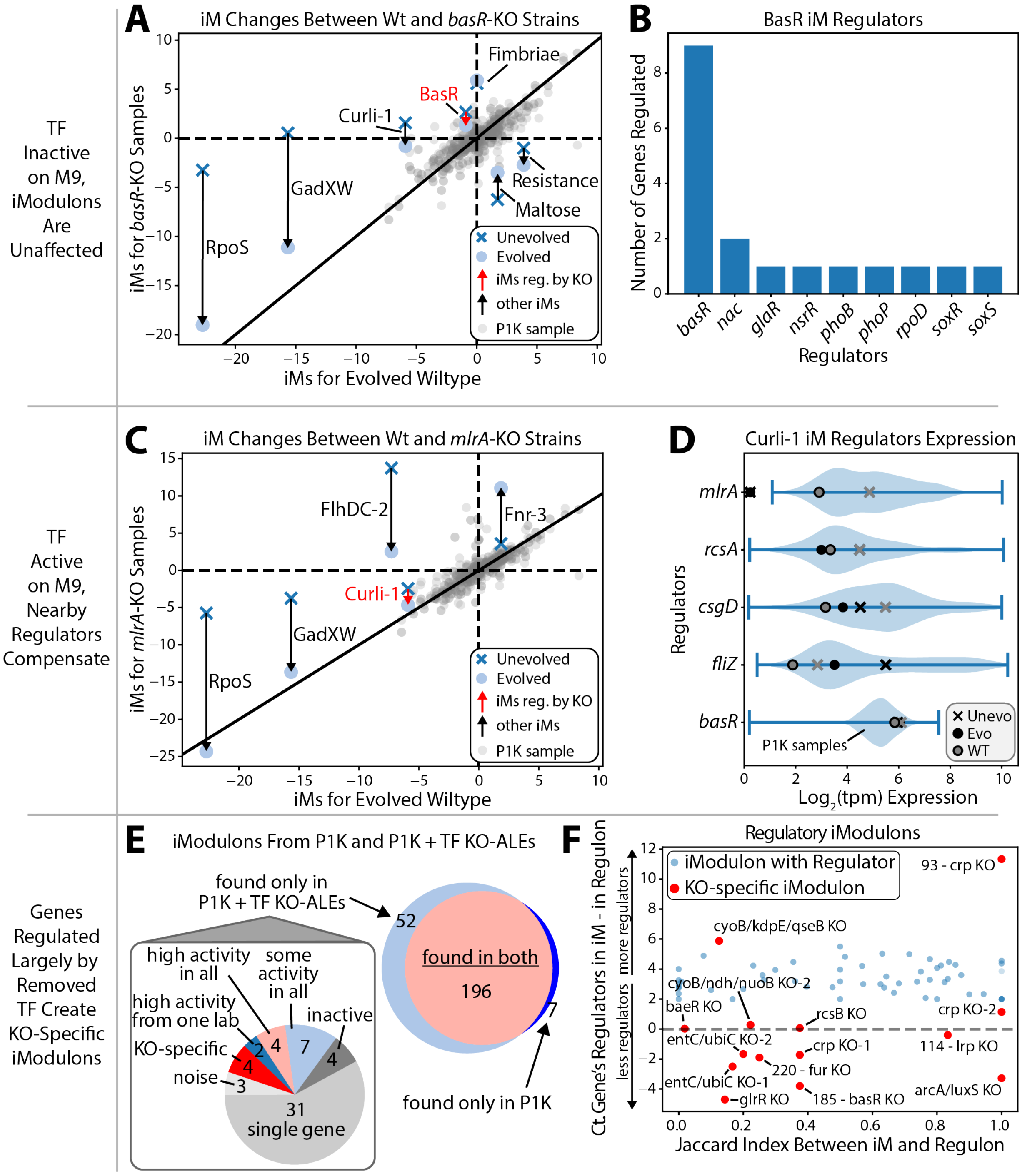

Figure 3.2: **(A)** Relative iModulon changes for all measured *basR* KO samples. **(B)** The number of genes regulated by each regulator of the BasR iModulon. **(C)** Relative iModulon changes for all measured *mlrA* KO samples. **(D)** Differential expression of the regulators of Curli-1. **(E)** The iModulon pipeline was rerun with P1K and samples from this study. **(F)** All iModulons with regulators are shown alongside the KO-specific iModulons of both PRECISE1K and this study.

iModulons.

3.2.5 ALE restores growth-important gene expression when deleted TFs are active with large regulons

Crp (cyclic AMP receptor protein) regulates a wide variety of genes mostly encoding enzymes involved in carbon metabolism and transport [111] and has long been the subject of intense interest in microbiology. *Crp* has the largest number of directly regulated genes [107], but through evolution its KO can be adjusted for. This is possibly due to the fact that *crp* has relatively few targets of which it is the only regulator and therefore other regulators with overlapping targets are able to maintain normal expression of most growth-important genes.

Figure 3.3A shows the iModulon changes for both the *crp* KO and its evolutions. There are two *crp* iModulons (ignoring KO-specific iModulons) in P1K whose activities are closely connected (Figure 3.3B): Crp-1, which contains a large number of genes primarily involved in carbon metabolism and Crp-2, which is dominated by the *gatYZABCD* operon (a transporter of galactitol [112]). Crp-2 is substantially downregulated in both the unevolved and the evolved *crp* KO strains while Crp-1, which is inactive on M9 (see Supplemental Figure 4), is only minorly affected by the KO or evolution. This may be due to or enabled by the fact that Crp-1 and Crp-2 largely correspond to class I and class II *crp* binding respectively [106], thus showing that *crp* class II binding appears to be more active on minimal glucose media than *crp* class I binding.

The evolution does not restore normal expression of *crp*'s regulon on the whole (as measured by the Crp iModulons), but rather restores the expression of the most growth important genes such as *ptsG*. Figure 3.3C shows the most differentially expressed genes regulated by *crp* in the *crp* KO samples, which show evolution's ability to rebalance the most growth-important

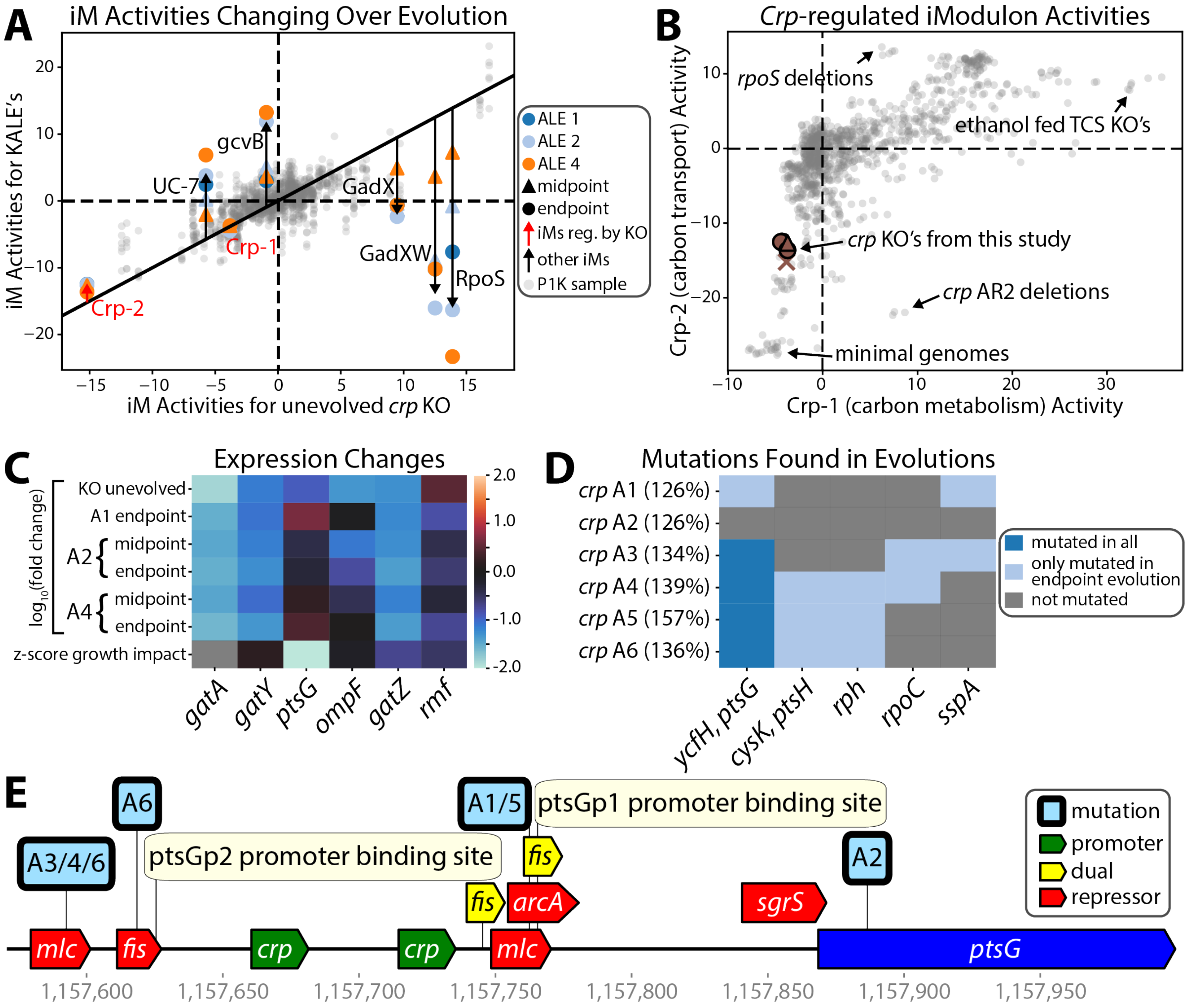

Figure 3.3: **(A)** Relative iModulon changes for all measured samples. iModulons that are differentially expressed are labeled. The relatively small amount of change to the *crp* iModulons over evolution shows that evolution does not restore the whole TRN but rather the genes with strong growth impacts. **(B)** Comparison of Crp-1 and Crp-2 iModulon activities across PRECISE1K. Outliers are annotated. **(C)** Only genes with differential expression in either the unevolved or evolved samples are shown. The most notable change is the restoration of *ptsG* to normal expression. All expression values are relative to the unevolved wildtype samples. The bottom row shows the impact to growth rate for the removal of each gene. The expression profile of the midpoint sample for A1 is contaminated and is thus removed. **(D)** All mutations that are found in at least two evolved strains are shown. There is a clear evolutionary selection pressure for *ptsG* promoter mutations. The percentages are the growth rate increase relative to the unevolved *crp* KO strain (i.e., crp A1 grows at a 2.36-fold higher rate). **(E)** The promoter and repressor sites for *ptsG* and the location of the mutations for each of the sequenced endpoint strains.

target of *crp* after its KO, *ptsG*. The *crp* evolution from our study shows similar results to a different *crp* KO-ALE study [100], both of which contain convergent mutations to *ptsG* repressor sites. The mutations found in the evolutions can be seen in Figure 3.3D with convergent mutations upstream of *ptsG*, a vital gene for glucose import. These mutations, visualized in Figure 3.3E, target repressor binding sites of other regulators.

Crp normally promotes *ptsG* and the KO thus severely downregulates the expression of *ptsG*, but the evolved samples are able to restore normal *ptsG* expression by reducing repressor activity. A2 is the only evolution without a *ptsG* mutation on a repressor site and instead has a mutation shortly downstream of the *ptsG* start codon. A2 is the slowest grower but this mutation does restore normal expression of *ptsG*, possibly through inhibiting *sgrS*, an sRNA which inhibits *ptsG* expression and binds near the start codon [113]. These mutations are the first to arise in the strains, even before common M9 evolution adaptations such as RNAP mutations. This further implies there is a very strong selection pressure to restore normal *ptsG* expression within these evolutions.

Fur primarily acts as a regulator of iron transport/utilization and has a large regulon of 132 genes [107], but the majority of the genes it regulates are also regulated by other TFs. There are two *fur* iModulons in P1K whose activities are highly coordinated with each other (Figure 3.4A): Fur-1 which is dominated by the *entCEBAH* operon which helps synthesize enterobactin to provide iron for metabolic pathways [116] and Fur-2 which consists of ABC iron transport proteins. Similar to the *crp* evolutions, the *fur* KO and evolution drastically changes the activity of Fur-1 while leaving Fur-2, which is less active on M9 (see Supplemental Figure B.4), relatively unchanged. Instead of restoring normal expression of the *fur* iModulons (Figure 3.4B), the evolved strains restore normal expression of high growth-impact genes, most notably

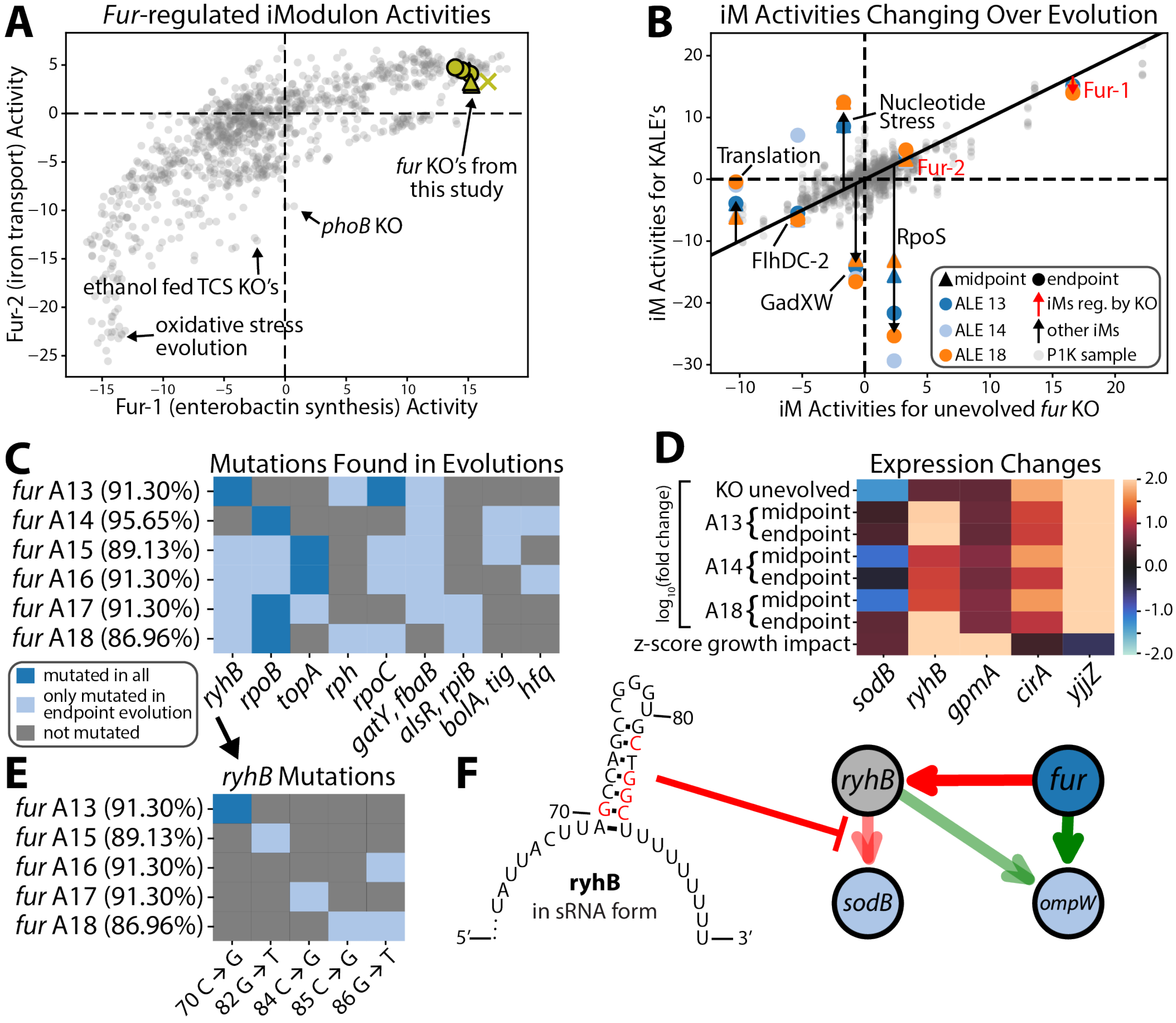

Figure 3.4: **(A)** Comparison of Fur-1 and Fur-2 iModulon activities across PRECISE1K. Outliers are annotated. **(B)** The differentially expressed iModulons for the evolved samples show that the majority of the evolution focuses on common growth-promoting adaptations rather than restoring the TRN to the normal state. **(C)** Mutations found in at least two independent lineages are visualized and the percentages are the growth rate increases relative to the unevolved *fur* KO strain. **(D)** Only genes with differential expression in either the unevolved or evolved samples are shown. All expression values are relative to the unevolved wildtype samples. The bottom row shows the impact to growth rate for the removal of each gene. **(E)** The specific sequence changes for the strains with an *ryhB* mutation are shown. **(F)** The structural location of the *ryhB* mutations on the sRNA form of *ryhB* [114], with red letters representing each mutated position. These mutations have been shown to reduce *ryhB*'s ability to regulate its targets [115].

sodB through mutations to *ryhB*.

RyhB is a sRNA which regulates many of the same targets that *fur* does and is normally repressed by *fur* [117]. In the unevolved *fur* KO, *ryhB* is unrepressed and thus highly expressed which in turn severely represses *sodB*, a superoxide dismutase (Figure 3.4D). The evolution has convergent mutations to some common ALE targets, such as RNAP and *topA* (Figure 3.4C), but also mutates a specific region of *ryhB* (Figure 3.4E). This region of *ryhB* is known to play an important role in *sodB* regulation and changes to it have been shown to reduce its ability to repress *sodB* [115]. A13 is the only lineage with such a mutation in its midpoint evolution, which is shown to be the only midpoint evolution with normal expression of *sodB*. Interestingly A14 recovered growth with only a 5.3% representation of this mutation in its endpoint evolution population sample, showing there are possibly multiple potential adaptation strategies to these KOs.

While these *ryhB* mutations are convergent among the independent lineages, unlike the *ptsG* mutations found in the *crp* KO samples, they are not commonly found in the midpoint samples and instead RNAP mutations are more often the first to become dominant. This further shows the high potential growth impact of RNAP mutations, which has been well documented [118].

3.2.6 *Lrp* KO-ALE does not recover growth and is associated with a large interconnected regulatory network

Lrp (leucine-responsive regulatory protein) is a global TF which coordinates cellular metabolism functions with the nutritional state of the cell [119]. *Lrp* notably regulates amino acid anabolism and catabolism [120], stationary phase adaptations [121], and nutrient transport [122] among many other metabolic processes.

Figure 3.5A shows a variety of measures about the TRN which make it clear that *lrp* is an outlier in many ways. It has the second largest set of genes for which it is the only known regulator, its network is highly complex (inferred by the high skewness of its subnetwork), and its high betweenness centrality shows how a majority of the pathways between regulators pass through it. Figure 3.5B shows a directed network of regulatorily interconnected TFs which includes 74.4% of all regulators. In this network, 28.9% of all TF-TF pairs are connected through regulation. If *lrp* is removed from this network, this percentage drops by 54%. For reference, if *nac* is removed this drops by 41.2%, *crp* drops by 22.5%, and *fur* drops to 12.1%. *Lrp* is in a unique position in the TRN as one of if not the only regulator that if removed severely disrupts the network structure of TF-TF regulation.

The *lrp* KO-ALE strains stand alone among our KOs as being unable to recover growth. There are no TF-specific convergent mutations among its evolved lineages and what mutations do exist are limited to well studied growth-promoting ones such as RNAP mutations [109] (see Supplemental Figure B.5). Transcriptional changes are largely limited to a downregulation of stress-related genes and slight upregulation of ribosomal subunits, which are common across ALEs [109]. The growth rate does improve through evolution, although this evolution contains no adaptations specific to the KO itself.

Lrp regulates numerous operons, many of whose expression is largely changed through *lrp*'s KO but unchanged by the subsequent evolution. Some of these operons are highly growth important, such as *gltBDF* and *livKHMGF* which the KO severely downregulated (Figure 5C). The evolution was unable to rebalance these operons to return the transcriptome to a healthy state (Figure 3.5D), thus limiting its growth capabilities.

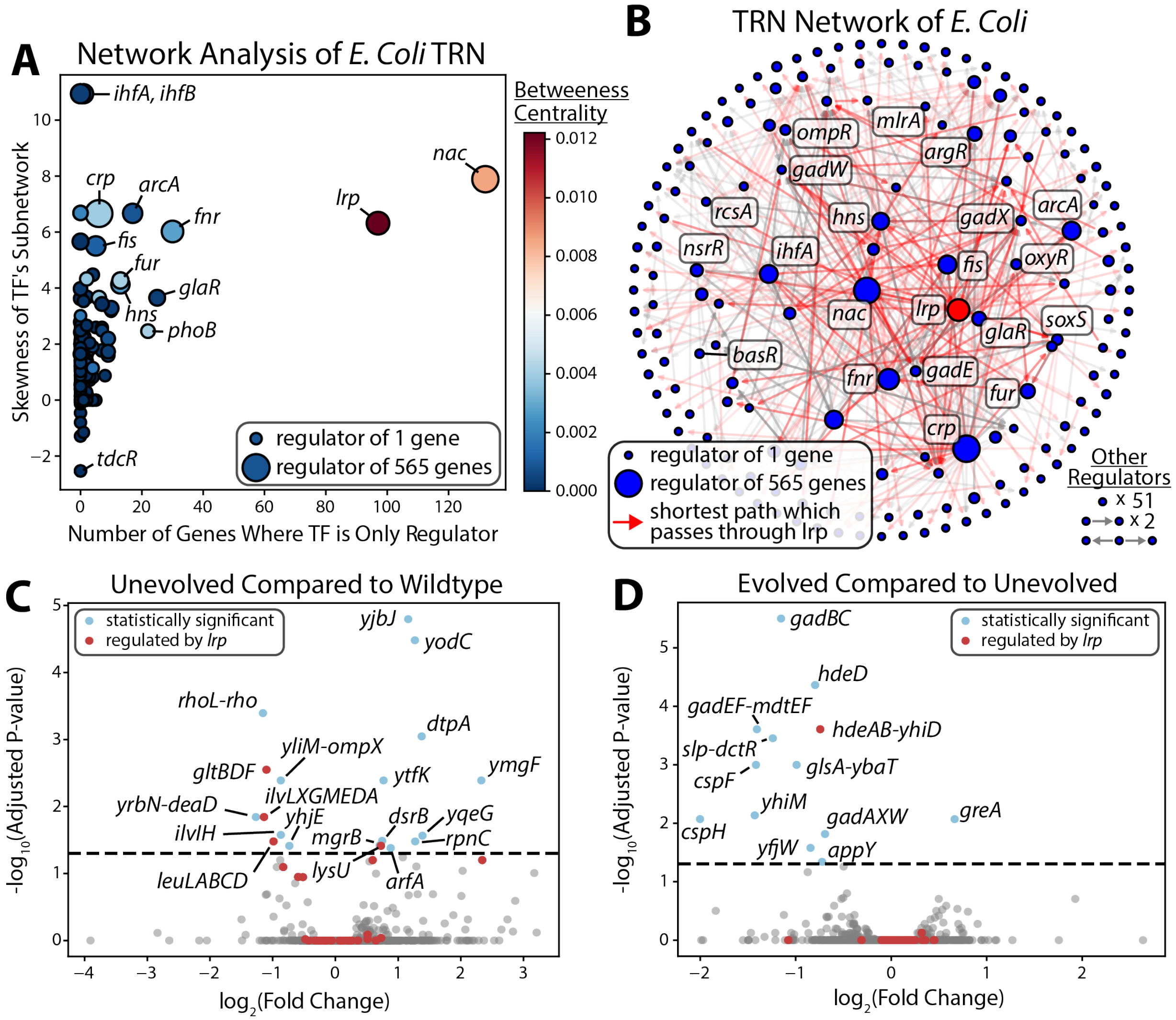

Figure 3.5: **(A)** Skewness indicates the skew of the distribution of the number of degrees for each TF's regulon network, higher skew meaning more complexity. Betweenness centrality is a measure of how many other shortest paths between the regulators pass through a certain regulator. *Lrp* along with *nac* are outliers in all of these measures. **(B)** The TRN network of *E. coli* according to RegulonDB [107]. Only genes that act as regulators are shown and regulatory networks unconnected to this central one are summarized in the bottom right (25.6% of regulators). While many genes are highly connected in the network visualized, *lrp* is a central element. All shortest paths between any two regulators that pass through *lrp* are colored red (60.4% of total shortest paths). **(C)** A DEO (differentially expressed operons) plot comparing the unevolved *lrp* KO to the wildtype, showing the large effect of the KO. **(D)** A DEO plot comparing the fastest endpoint strain (A23) of the *lrp* KO-ALE to the unevolved *lrp* KO strain. The evolution is not able to modify the expression of the large majority of *lrp*'s regulon.

3.2.7 Certain TF KOs and subsequent evolutions affect readiness of TF-associated substrates

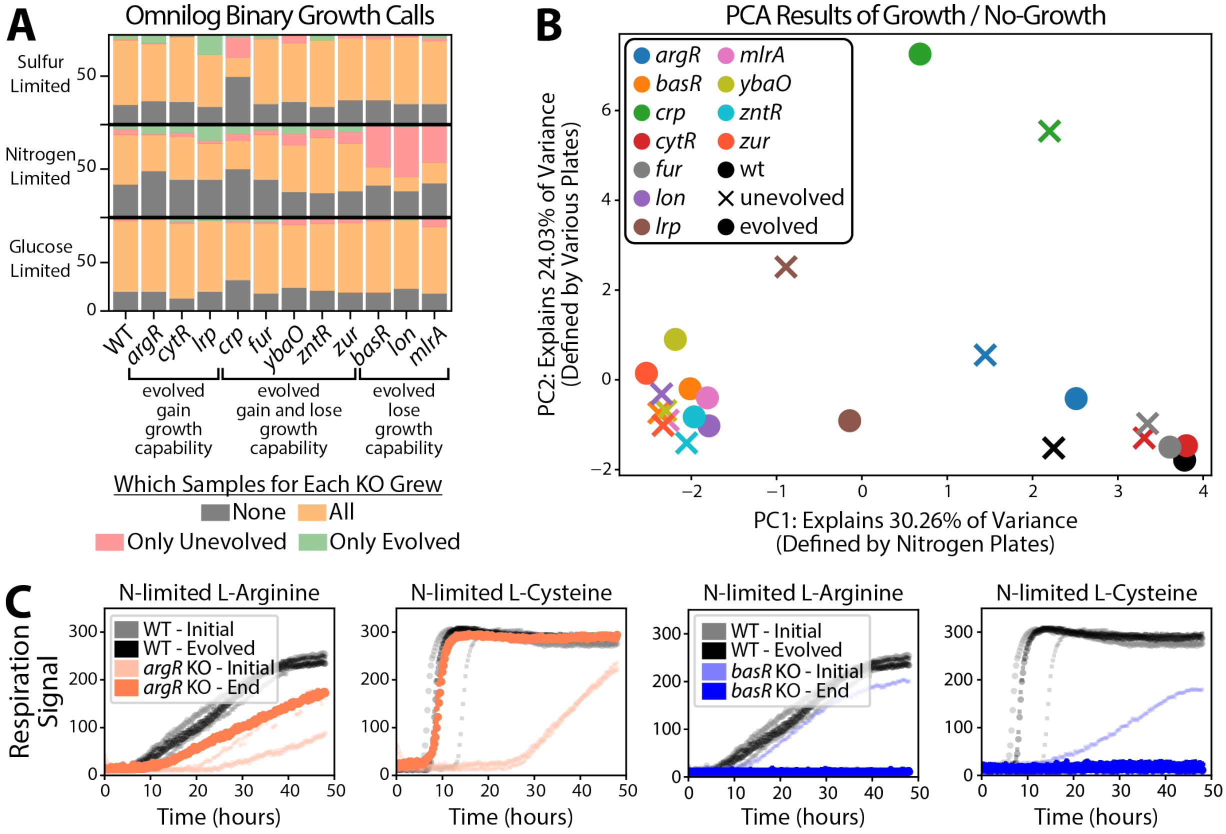

Figure 3.6: **(A)** The substrate readiness results for all of the KO-ALEs, showing that evolution overall reduces the ability of a strain to grow on new substrates as it focuses on increasing growth for the substrate it evolved on. See Methods: Biolog Plates and Analysis for more information. **(B)** A principal components analysis carried out on the binary growth / no-growth calls of the phenotyping plates results. The top two highest variance-explaining principal components are shown. **(C)** A selection of the growth profiles from the substrate readiness plates, showing the growth capability differences between the *argR* and *basR* KO strains on nitrogen-limited amino acid plates.

Despite the lack of convergent mutations or KO-specific transcriptional changes among many of the KO-ALE strains, the substrate readiness plates show distinct phenotypic changes between the wildtype, unevolved KO, and evolved KO strains. The plates showed that most of the evolved strains grew on less substrates than their unevolved ancestors as they specialized towards growth on minimal media supplemented with glucose (Figure 3.6A). The strains that largely gain growth capabilities (*argR*, *cytR*, and *lrp* KOs) do not have TF-specific mutations. Many strains show a reduction in ability to grow in nitrogen-limited conditions over the course of their evolution (the *lon* KO evolved strain loses the ability to grow on 74.63% of nitrogen-limited conditions, *basR* 69.84%, *mlrA* 63.93%, *zur* 8.06%, *zntR* 7.46%, *ybaO* 4.92%, WT 3.45%), presumably a consequence of their evolution on nitrogen rich M9 media. It should be noted, however, that these growth losses seen largely in the *basR*, *lon*, and *mlrA* KO strains may also be experimental artifacts.

A principal component analysis of the binarized phenotype data indicates that the highest variance amongst the strains is explained by differences in nitrogen substrate utilization (Figure 3.6B). The second highest variance explanatory component represents variance in carbon, nitrogen, and sulfur utilization. The *argR*, *crp*, and *lrp* KO-ALE strains exhibit the largest shifts in substrate utilization. While the changes for *crp* and *lrp* KO-ALEs were previously discussed, this large shift in substrate utilization is unexpected as the *argR* KO strains showed no convergent mutations and quickly recovered growth.

The unevolved and evolved KO strains for *argR*, a dual-regulator of arginine import and biosynthesis [2], modify its growth capability on numerous amino acids. The unevolved *argR* KO strain loses the ability to grow on five amino acids under nitrogen-limited conditions that the wildtype can grow on (arginine, asparagine, cysteine, glutamic acid, and lysine). The evolved *argR* KO strain restores growth on four of these including arginine and increases growth rates on many of the other amino acids (see Supplemental Figure B.6). [123]. Contrasting with the *argR* case, the fastest *basR* KO-ALE endpoint strain does not gain the ability to grow on any new substrates and instead loses the ability to grow on nearly any amino acid (Supplemental Figure

B.7). A few of these nitrogen-limited amino acid cases can be seen in Figure 3.6C.

Some strains such as the *cytR* KO show few differences between the unevolved and evolved strains. The *cytR* KO strains show no change compared to the wildtype samples on any of the six cytidine conditions, which modulates *cytR* activity [124]. *YbaO* (*decR*), which is thought to play a role in cysteine detoxification [125], showed distinct differences on nitrogen-limited cysteine conditions but relatively little difference caused by the evolution. Similar to what we found for some of the TFs that are active on minimal glucose media, it appears that activity alone is not enough to determine if a KO will lead to a detrimental effect on growth.

3.3 Discussion

TF KO-ALEs reveal a wide variety of transcriptional, sequence, and phenotypic adaptations which are highly dependent on the activity of the TF on the growth medium and the size of the TF's regulon. Most of the TF KO strains were able to recover growth rates through non TF-specific evolutionary strategies. Some of these general recovery strains, despite an initial growth defect, are not active on minimal glucose media and the cell was able to restore growth without large changes to its TRN. Others, such as the *mlrA* KO-ALE strains, recovered growth through the utilization of regulators with overlapping targets to the removed TF without the requirement of TF-related mutations. The *crp* and *fur* KO evolved strains recovered growth both through common minimal media adaptations and by restoring the expression of highly growth-important genes through convergent mutations to elements of their own TRNs, while leaving the majority of their regulons highly differentially expressed. The *lrp* KO strains stand alone as being unable to recover growth, which is likely a consequence of *lrp*'s unique central position within the TRN and large number of genes for which it is the only regulator. The differentially expressed genes of

TF KOs have been shown to vary between 0 and 63% related to the removed TF [126], showing the wide range of potential impacts of TF KOs.

E. coli can restore growth through evolution without KO-specific mutations for most of the TFs included in this study. Often this is because the TF is not normally active on M9, but the highly connected nature of the TRN means that even for many active TF KOs, their regulons are still regulated by other TFs and the strains quickly recover the growth loss caused by the KO. The changes in expression or sequence through the evolution of these TF KOs are not related to the removed TF's regulon and instead represent evolutionary adjustments to M9 supplemented with glucose seen across a wide variety of ALEs [109]. An evolved 24% reduced genome derivative of MG1655 grew only 13% slower than evolved wildtype whereas the unevolved reduced genome strain grew 69% slower [127], inferring that the unexplained initial growth defect seen in many of our study's strains may be more general KO-response rather than a TF KO-specific response.

Some TF KO-ALEs, such as *crp* and *fur*, are able to recover growth through convergent mutations. *Crp* KO-ALE strains restore normal expression of *ptsG* through targeted mutations to repressor binding sites upstream of *ptsG* (in agreement with another study [100]) and *fur* KO-ALE strains mutate a specific region of *ryhB* which helps rebalance *sodB* expression to standard levels [115]. Both the *crp* and *fur* KO-ALE strains, however, do not return their respective regulons to normal levels but rather restore the expression of the few most growth-important genes. A *pdhR* KO-ALE on glucose minimal media resulted in convergent mutations to elements of its TRN [101] similar to the *crp* and *fur* KO-ALEs from this study. *PdhR* has a relatively small regulon of 53 genes but contained in these are highly growth-important genes such as the pyruvate dehydrogenase complex [101]. *PdhR* has no clear iModulon in order to infer its activity on said media but presumably is active and could thus be categorized alongside the *crp* and *fur*

KOs.

Lrp KO-ALE is not able to restore normal growth. This may be due to a handful of unique attributes of *lrp*, but most notably its large regulon of approximately one third of the genome [128], the large number of genes of which it is the only regulator, and its central location within the TRN. No other gene in *E. coli* plays a more central role in connecting different TFs to each other through regulation. Evolution is able to improve the growth rate of the *lrp* KO, but primarily through generic M9 growth-promoting mutations and expression changes that are seen across many samples from this study and numerous other ALE studies [109]. In a whole cell network study connecting together the TRN and metabolic networks, the Lrp-leucine complex was one of the central connections between the two networks, thus showing its central importance to not just regulating TFs but also its influence over metabolism [129]. The regulon of *lrp* remains largely unrestored through evolution, including growth important operons such as *gltBDF* and *livKHMGF*.

Our characterization of substrate readiness showed a decreased ability of the evolved strains to grow on other substrates, so there are likely other conditions where these KOs may necessitate KO-specific adaptations. A study about novel ppGpp function found unexpected substrate readiness differences following gene knockouts [130], giving some evidence that our observed growth differences seemingly unrelated to the removed TF may actually be the result of said TF's removal and not an experimental artifact. A large-scale study of *E. coli* phenotyping plates found the most no-growth calls on carbon sources and large disagreement of growth/no-growth calls on nitrogen sources [131]. That said, despite a lack of clear evolutionary adjustments, the evolution of TF KOs did largely modify which conditions the strains could grow on, sometimes changing growth behavior on substrates related to the removed TF.

TF KO-ALEs teach us a similar lesson to the metabolic KO-ALEs in that both create bottlenecks, in the TRN and in the metabolic map respectively, that are overcome using nearby existing connections and genetic mutations within these networks. A *nac* KO-ALE serves as a potential follow-up study to this one, as it, like *lrp*, regulates a large number of genes many of which have only it as a regulator and connects many TFs to each other in the TRN. Similar to how some of our removed TRN connections enabled new growth capabilities, another study showed how creating new regulatory links can also confer a growth benefit [132]. Additionally, the modification as opposed to removal of central regulators has also been shown to modify phenotype [133]. Despite large initial growth impacts, it appears that the majority of TFs can be removed from *E. coli* on M9 minimal media and growth will quickly recover as the high connectivity of the TRN leaves it with few vulnerabilities. These results reinforce the long-standing view of the TRN as a highly adaptable network and begin to systematically uncover mechanisms by which this adaptability is achieved.

3.4 Methods

Strain Information

All strains were selected from the Keio collection [134]. Round 1 strains (*argR*, *crp*, *cytR*, *fur*, and *lrp*) were evolved at the Center for Biosustainability at Denmark Technical University while round 2 strains (wt, *basR*, *lon*, *mlrA*, *ybaO*, *zntR*, and *zur*) were evolved at University of California San Diego.

ALE and Growth Characterization

ALE was performed using 6 independent replicates of each TF KO. All ALE experiments were conducted at 37°C with a stirring speed of 1100 rpm for proper aeration. Each individual experiment was passed to a new culturing flask around an OD600 nm = 0.6. Cultures were always maintained in excess nutrient conditions assessed by non-tapering exponential growth. The evolution was performed for a sufficient time interval to allow the cells to reach their fitness plateau. The growth medium for all samples was M9 minimal medium with 4 g/L glucose, supplemented with Wolfe's vitamin solution and trace elements.

Samples are named in the following convention: TF KO'd, A(LE) number, F(lask) number, I(solate) number, R(eplicate) number. For example - argR A1 F14 I1 R1 is the A1 independent lineage of the *argR* knockout strain and is the first replicate from the first isolate taken from the 14th flask. Later flask numbers indicate longer evolution times. All knockouts also have an A0 strain, which is the unevolved sample. Throughout the paper, endpoint flask refers to the higher flask number from a lineage while the midpoint flask refers to the other non-zero flask number from the lineage.

DNA-sequencing

A clone from the midpoint and endpoints of the evolved strains was picked for DNA sequencing. The strains were grown in an M9 minimal medium supplemented with 4 g/L glucose. Total DNA was sampled from an overnight grown culture at an OD600 nm = 0.6. Nucleic acid isolation, library preparation, and subsequent analysis were performed as previously described [135]. Briefly, genomic DNA was isolated using a Nucleospin Tissue kit including treatment with RNase A. Resequencing libraries were prepared following the manufacturer's protocol using

Nextera XT kit. Sequencing was performed on an Illumina HiSeq. Sequence data is available at https://aledb.org/ale/project/138/.

RNA-sequencing

All samples from both rounds were prepared and collected in biological duplicates at UCSD. 3 ml of culture sampled from an overnight grown culture at an OD600 nm = 0.5 was added to 6 ml of Qiagen RNA-protect Bacteria Reagent after sample collection. This solution was then vortexed for 5 seconds, incubated at room temperature for 5 minutes, and then centrifuged. The supernatant was then removed and the cell pellet was stored at -80° C. The Zymo Research Quick-RNA MicroPrep Kit was used to extract RNA from the cell pellets per vendor protocol. On-columns DNase treatment was performed for 30 minutes at room temperature. Anti-rRNA DNA Oligo mix and Hybridase Thermostable RNase H [88] was used to remove ribosomal RNA. Sequencing libraries were created using a Kapa Biosystems RNA HyperPrep per vendor protocol. RNA-sequencing reads were processed using https://github.com/avsastry/modulome-workflow. Data is available at NCBI GEO GSE266148.

iModulon Computation

RNA-sequencing data was used to create iModulon activity levels of our strains using PyModulon [37] which is available at https://github.com/SBRG/pymodulon. Activities of iModulons were compared to samples from PRECISE1K which is easily accessible using iModulonDB [5]. The calculations of new iModulons for our dataset in addition to PRECISE1K was performed using modulome-workflow [110] which is available at https://github.com/avsastry/modulome-workflow. See https://imodulondb.org/ [5] for a more complete description of iModulons and their calculation.

Basal iModulon Transformation

In the final generation of the M and A matrices within our workflow, the sign of specific components in M was inverted to ensure a predominantly positive distribution of gene weights. This adjustment involved reversing the sign of the corresponding columns in the M matrix and the associated rows in the A matrix. To improve the interpretability of A matrix, adjustments were made such that higher values correspond to an increased regulatory activity (basal activity). Each iModulon was assigned a direction based on the function of its canonical regulator and, when available, the activity observed in knockout (KO) samples. To align the A matrix values with a baseline, the values were shifted such that the minimum approximates zero. This was achieved by subtracting the 95th quantile for iModulons with a positive direction, and the 5th quantile for those with a negative direction, ensuring minimal influence from outliers. If the direction was positive, the signs of both matrices A and M were subsequently flipped to maintain consistency.

Biolog Plates and Analysis

The OmniLog system was used to generate the media screens. First overnight cultures were grown in 4 mL of M9 4g/L glucose medium at 37° C with shaking. Pellets were collected by centrifuge and pellets for PM01 plates were washed twice by M9 no carbon medium while pellets for PM03B and PM04A plates were washed twice using IF0a medium (supplied by OmniLog). 42%T and 85%T sample solutions were prepared in M9 no carbon medium with 1X DyeA (supplied by OmniLog) for PM01 plates. 42%T and 85% sample solutions were prepared in IF0a medium with 1X Carbon (Na-succinate and FE-Citrate) and 1X DyeA. 100 μ L of 85% T sample solution was placed in each well. The plates were run on the OmniLog machine at 37° C for 48 hours. Opacity readings were generated of the plates over these 48 hours.

In order to convert the opacity respiration readings to growth curves, the signal from every well is processed through a Savitzky-Golay filter to smoothen the data. A window length of 50 and polynomial degree of 3 is used for said filter. The maximum signal value is recorded and a control group is formed of the negative control wells. A 1 sided z-test is performed to calculate p-values associated with each well and are corrected for multiple hypothesis tests using the Bonferroni correction. If the adjusted p-value is below 0.05 the well is considered to have a significant growth signal and is assigned to have grown, else no growth is assumed.

Acknowledgements

This work was supported by The Novo Nordisk Foundation (NNF) Center for Biosustainability (CfB) at the Technical University of Denmark (NNF20CC0035580).

Chapter 3 in part is a reprint of material submitted for publication in *Proceedings of the National Academy of Sciences (PNAS)*:

- **C Dalldorf**, Y Hefner, R Szubin, J Johnsen, E Mohamed, G Li, J Krishnan, AM Feist, BO Palsson, DC Zielinski. 2024. “Diversity of transcription regulatory adaptation in *E. coli*” The dissertation author was the primary author.

Chapter 4

Data-driven modeling of bacterial transcriptional regulation

The growth of bacterial gene expression datasets has offered unprecedented coverage of achievable transcriptomes, reflecting diverse activity states of the transcription regulatory network. Machine learning methods like Independent Component Analysis (ICA) can decompose gene expression datasets into regulatory modules and condition-specific regulator activities. Here, we present a workflow to utilize inferred regulator activities to construct quantitative models of promoter regulation in *E. coli.* Resulting models are validated by predicting condition-specific TF effector concentrations and binding site motif strength based on differential gene expression data alone. We show how reconstructed promoter models can capture multi-scale regulation and disentangle regulator interactions, including resolving the apparent paradox where argR expression is positively correlated with its regulon despite being a repressor. We applied the workflow for all regulator-linked components extracted by ICA, demonstrating the scalability of the workflow to capture the *E. coli* TRN. This work suggests a path toward systematic, quantitative reconstruction of transcription regulatory networks driven by the large-scale databases that are now available for many organisms.

4.1 Background

The transcription regulatory network of bacteria is a critical determinant of cell state and responsiveness [136], and largely involves trans-acting regulatory proteins which bind to promoter regions to promote or deter recruitment of RNAP. Binding sites and activities of key regulators such as transcription and sigma factors have been painstakingly determined over the years through experiments such as ChIP and gSelex [137]. Meanwhile, observations of transcription regulator states, in the form of gene expression datasets, have become increasingly available for diverse conditions. In *E. coli*, the scale of both TF binding identification and RNA-seq have been rapidly increasing, with only 121 of the estimated 300 TFs having any known gene targets in 2003 [138] compared to 232 TFs with gene targets with strong or confirmed confidence today [137]. The development of sequence analysis tools has enabled binding site identification purely through computational means [137], including some tools which can additionally predict binding strength [139]. As of 2024, NCBI GEO contains 23,890 individual samples of RNA-seq for *E. coli*, while at the end of 2010 there were only 4,127 samples [140]. This abundance of data suggests that there exists the potential for an integrative understanding of transcription regulatory network function in *E. coli*.

TRN network inference from gene expression datasets is a classic problem in systems biology [141]. Approaches to infer transcriptional regulation from large datasets generally focus on identifying regulatory interactions and strengths rather than explicitly modeling the biochemical nature of the resulting regulation. These methods for TRN estimation contrast with another paradigm for network assembly, network reconstruction, which focuses on the bottom-up structured assembly of biological knowledge through manual curation [142]. Although transformative in modeling metabolism, the reconstruction workflow when applied to transcription regulatory networks to date has relied on experimental mappings of transcription units [143] and computationally determined motifs [144] and resulted in largely qualitative (Boolean) [145] models. However, a key distinction of reconstructions compared to other knowledge bases such as encyclopedias [146] is that they are structured to enable the generation of quantitative models whose predictive performance can be evaluated. Thus, there is the potential for data-driven and reconstruction based approaches to be unified within a quantitative framework to more accurately capture transcription regulation.

Independent component analysis (ICA) has been a successful approach for *de novo* inference of transcriptional modules from gene expression databases [37]. The resulting activities of these components offer an estimate of the regulator activity on a given condition. Presumably, with a knowledge of transcription factor regulons and sufficient measurements of the outcomes of regulation, one can create a model to directly connect expression to biophysical measurements of TFs such as concentration and K_d values. This approach seeks to learn the function of TRNs from observing their behavior across conditions using methods like ICA which capture real regulon structure and activities, thereby mapping phenomenological parameters to mechanistic ones.

In this study, we develop a workflow to quantitatively reconstruct the TRN of *E. coli* by building mechanistic promoter models and parameterizing them utilizing large-scale gene expression data. We first define condition-specific regulon activities using ICA. We utilize known and inferred regulons to generate mechanistic models for all regulated promoters characterized to date. We then parameterize these models in a stepwise process, going from phenomenological ICA activities to fundamental biochemical parameters. We validate our approach by predicting effector metabolite concentrations and TF dissociation constants. The resulting models are used to disentangle complex regulation scenarios and interpret mutations. We extend this workflow to all regulons for *E. coli* with activities that can be inferred from ICA, representing a substantial fraction of all known regulation. This work lays the groundwork for a quantitative understanding of transcriptional regulatory networks in bacteria at a new scale.

4.2 Results

4.2.1 Constructing promoter models with data-inferred regulator activities

First, we describe the inference of condition-specific regulator activities from expression datasets. Various machine learning methods have been used to separate gene expression data into groups of genes forming co-regulated modules and regulator activities. Independent component analysis (ICA) has been demonstrated as one of the most effective methods based on the ability to capture experimentally-determined regulons. Independent components from gene expression analysis have been termed iModulons, or independently modulated groups of genes. The process of generating iModulons using ICA can be best understood as a blind-source separation of regulatory signals. ICA separates gene expression data (X) into groups of genes called iModulons (M) and the activity of these iModulons across the input samples (A) (Figure 4.1A). The latest version of iModulons produced for *E. coli* is the PRECISE1K dataset [106] which includes over 1000 individual expression profiles and generates 201 iModulons that altogether account for 86% of known regulatory interaction.

The premise of the modeling workflow is to use these iModulons as indicators of condition-specific regulatory activity, then subsequently infer a biophysical promoter model that matches both these regulator activities and resulting gene expression. The generation of this model involves inferring biological variables effector concentration, regulator concentration, and transcription K_d values, providing a basis for independent validation of the generated models. RegulonDB, a compendium of regulatory binding sites, provides a knowledge base of what genes are regulated by what transcription factors [137]. This information, combined with inferred regulator activity of said transcription factors from ICA enables us to create a biophysical model of transcriptional regulation. To do this, we created a thermodynamic model of regulated transcription with different promoter states represented along with binding and rate constants (Figure 4.1B). We also model effector concentrations along with their binding constants to transcription factors which require effectors.

The mathematical equations connecting these promoter states to each other and to the expression values are outlined in Methods: Mathematical Model Equations. In summary, they are able to calculate expression values for a specific gene based on cActivator and cInhibitor which, respectively, refer to the regulatory activity of a promoter and a repressor. The equations also require several biological constants to be set, such as $K_{d,RNAP}$, $K_{eq,opening}$, and $K_{d,Act-RNAP}$. These are set by calculating various possible solution sets that satisfy the underlying mathematical equations and picking the set that creates the best range of cActivator and cInhibitor values (see Methods: Selection of Biological Constants). Other values are hard set by the data itself, such as RNAP concentration, $K_{d,Act-Eff}$ and $K_{d,Inh-Eff}$ (see Methods: Selection of Biological Constants). Additional parameterization steps are further described in the methods.

Figure 4.1C gives a summary of the resulting workflow. With the mathematical model set and biological constants ready, we now need to generate the input cActivator and cInhibitor values. If a gene has only an inhibitor or activator, these can be directly solved as the number of equations and unknowns is equal. The ICA reconstructed expression value, M × A of the gene and the activating or inhibiting iModulon, is used as the expression value in the equation. If there is an inhibitor and activator for a gene, these are created using a genetic algorithm which creates a set of valid solutions to the expression equations and selects the solutions which best correlate with the activating and inhibiting iModulons. An additional greedy algorithm further optimizes these solution sets for input into the modeling software, GAMS (see Methods: Genetic and Greedy Algorithm Optimization). GAMS also requires the different activation types of the transcription factors involved, which can be a TF that is active without an effector, a TF that is active with a single effector molecule, or a TF that requires two copies of the same effector molecule to be active.

The process above is repeated for each gene contained in a specific activator-inhibitor pair we call a regulatory case. For example, all genes that are repressed by the Arginine iModulon and have no activating iModulon are grouped together as a regulatory case, which GAMS runs on independent of all other genes. The GAMS model optimizes for two criteria: 1) matching the input and output cActivator and cInhibitor values; and 2) matching the actual and predicted mRNA values. Additional details about the GAMS model are available in Methods: GAMS Model. The GAMS model outputs predicted mRNA values, GAMS optimized cActivator/cInhibitor values, and predicted biological constants for the regulators and promoter sites. Depending on the activation type of the associated transcription factors, these regulator-associated biological constants can include metabolite concentration, $K_{d,Act}$, $K_{d,Inh}$, and TF concentrations.

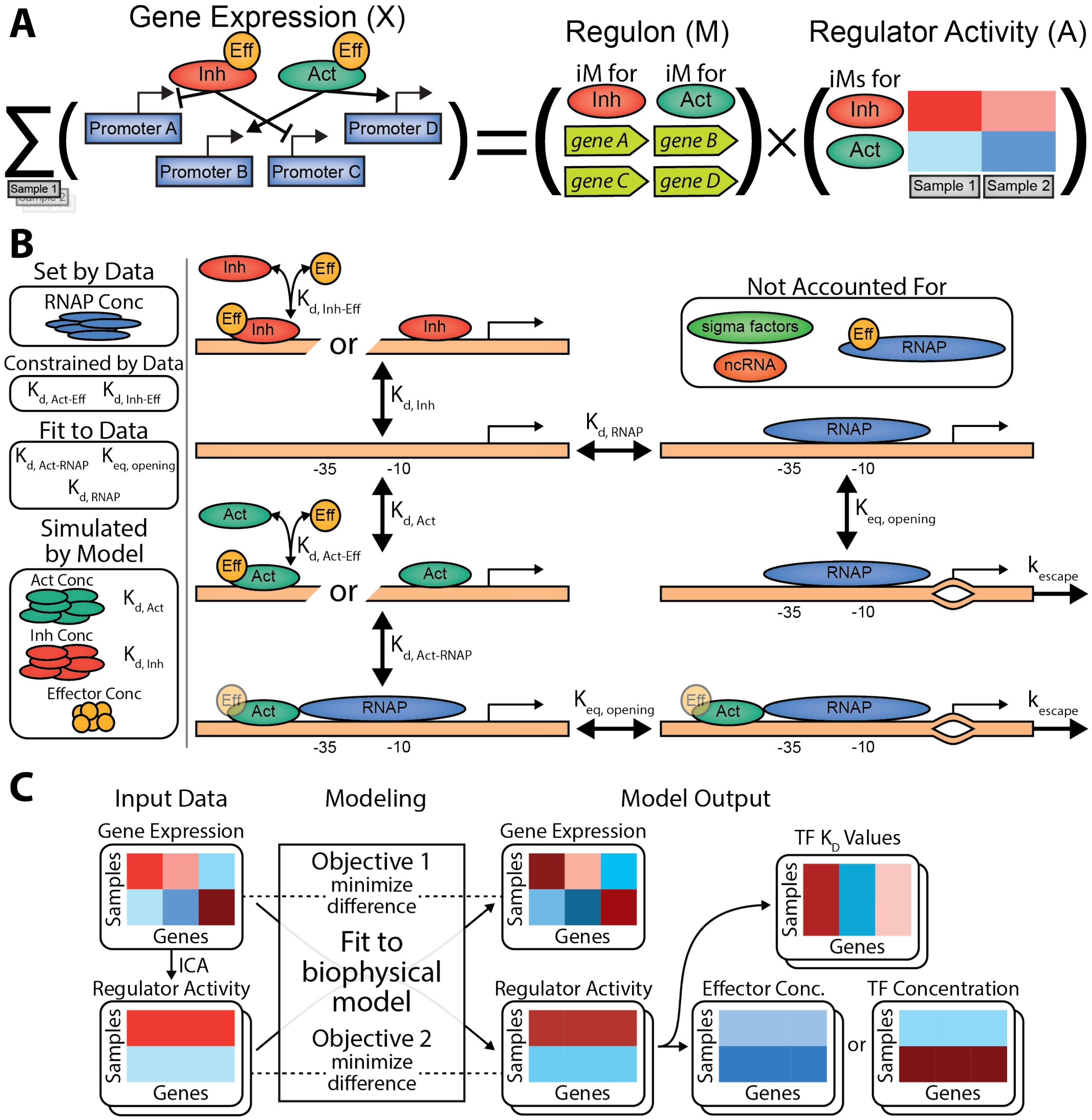

Figure 4.1: **(A)** Application of independent component analysis to expression data in order to approximate regulatory activity. **(B)** Mechanistic promoter model along with biological constants set by the data, fit to the data, and simulated by the model. The different promoter states this study's model incorporates are shown. Some relevant factors to gene expression are not modeled, such as sigma factors, non-coding RNA, direct RNAP binding effectors, and the concentration of genes. **(C)** Overall simplified workflow diagram for our model.

4.2.2 Multi-scale model resolves counter-intuitive regulation of arginine biosynthesis genes

We first examined the ability of the established transcriptional regulatory modeling framework to capture the regulation of the ArgR regulon, which regulates the arginine iModulon consisting of arginine biosynthesis genes. We observed that expression of *argA*, along with the other genes repressed by ArgR, is positively correlated with *argR* expression in the PRECISE1k database (Figure 4.2AB). This is counter-intuitive, as higher repressor expression logically should infer higher repressor activity and thus less expression. We also noticed that the Arginine iModulon has a much higher correlation to *argA* than *argR*. Thus, we hypothesized that the Arginine iModulon activity is a better approximate for effective ArgR activity than *argR* expression, and may actually have an inverse relationship with effective ArgR activity due to the effect of the ArgR activator arginine, which was an unknown in the model. Figure 4.2C shows a summarized view of *argR* and *argA* expression, highlighting the important role that arginine concentration itself plays in said regulatory network.

Accounting for both ArgR and arginine levels, our model is able to use this inferred regulatory activity from the iModulon to accurately recreate gene expression values (Figure 4.2D). This workflow is able to untangle this complex and non-intuitive regulatory network by incorporating arginine concentration. The predicted arginine concentrations highly correlates with experimentally measured arginine values [147] for a subset of conditions where these measurements were available (Figure 24.2E). This offers external validation of our workflow and highlights its ability to predict the metabolome using expression data.

In addition to validation by experimental data, we can also compare sequence-based predictions of *argR* binding strength for the modeled genes to their predicted K_d values. Higher binding strength would mean less of the active transcription factor would be necessary for expression, thus meaning binding strength and K_d values should be anti-correlated. We calculated the binding motif strength between *argR* and the promoter sites of the modeled genes (Figure 4.2F). This resulted in an expected highly negative correlation between predicted K_d values and sequence-based predictions of binding strength (Figure 4.2G).

4.2.3 Validation of models through prediction of effector concentrations

We next examined the PurR regulon, which regulates the Purine iModulon and has two different metabolite effectors, guanine and hypoxanthine [148]. We generated a model that accounts for both effectors interchangeably, due to unknown binding constants, and utilizes total purine as a variable (Figure 4.3A). For data from a joint expression and metabolite experiment [147], the resulting model is able to accurately predict this summed concentration at early time-points post glucose starvation. However, the model fails to predict purine concentrations during late starvation (Figure 4.3B). This failure could be due to a handful of factors, including the fact that the model does not yet account for sigma factors or direct RNAP effectors, as this data is primarily from stationary phase samples where *rpoS* and ppGpp play large roles in gene regulation [149].

Figure 4.3CD showcases how this model can accurately predict gene expression and recreates the input cInhibitor values for the majority of the genes. Similar to the Arginine case, the model's predicted K_d values also reliably anti-correlate with the sequence-based predictions of motif binding strength (Figure 4.3E).

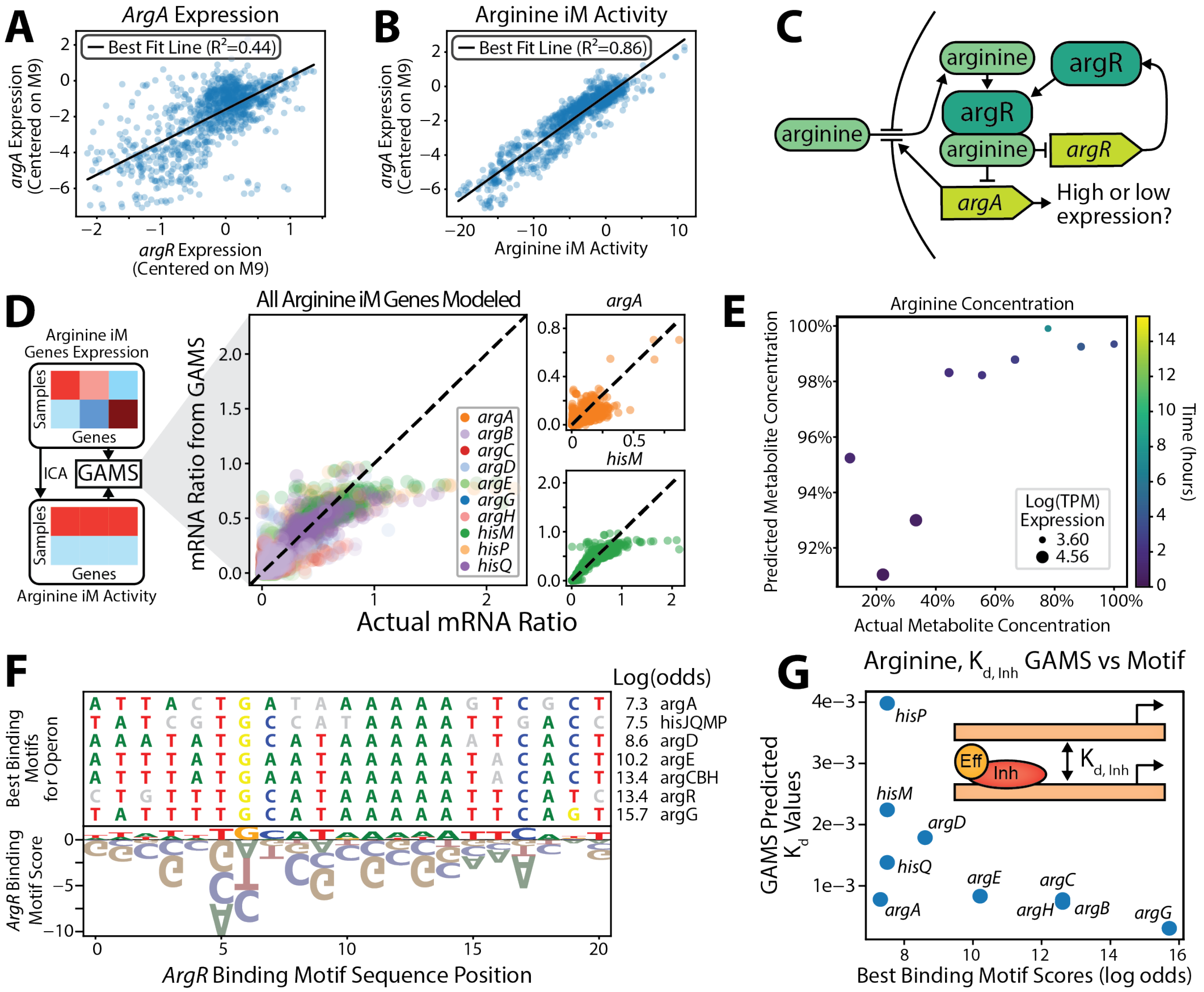

Figure 4.2: **(A)** *ArgA* expression is positively correlated with its repressor, *argR*. Samples are from PRECISE1K. **(B)** *ArgA* expression is highly positively correlated with the Arginine iModulon activity levels, showing it is a better indicator for regulatory activity than *argR*'s expression alone. **(C)** The regulatory network surrounding *argA* and *argR*, which is dependent on arginine concentration and also regulates various arginine processes including its import. **(D)** Inputs and outputs of GAMS model. GAMS is able to accurately predict gene expression. **(E)** Predicted and actual metabolite concentrations are highly correlated, which provides an external validation of the model. **(F)** *ArgR* binding motif alongside the most likely binding sites for the modeled operons. **(G)** Predicted GAMS K_d values for *argR* binding are highly correlated with the sequence-based binding calculations.

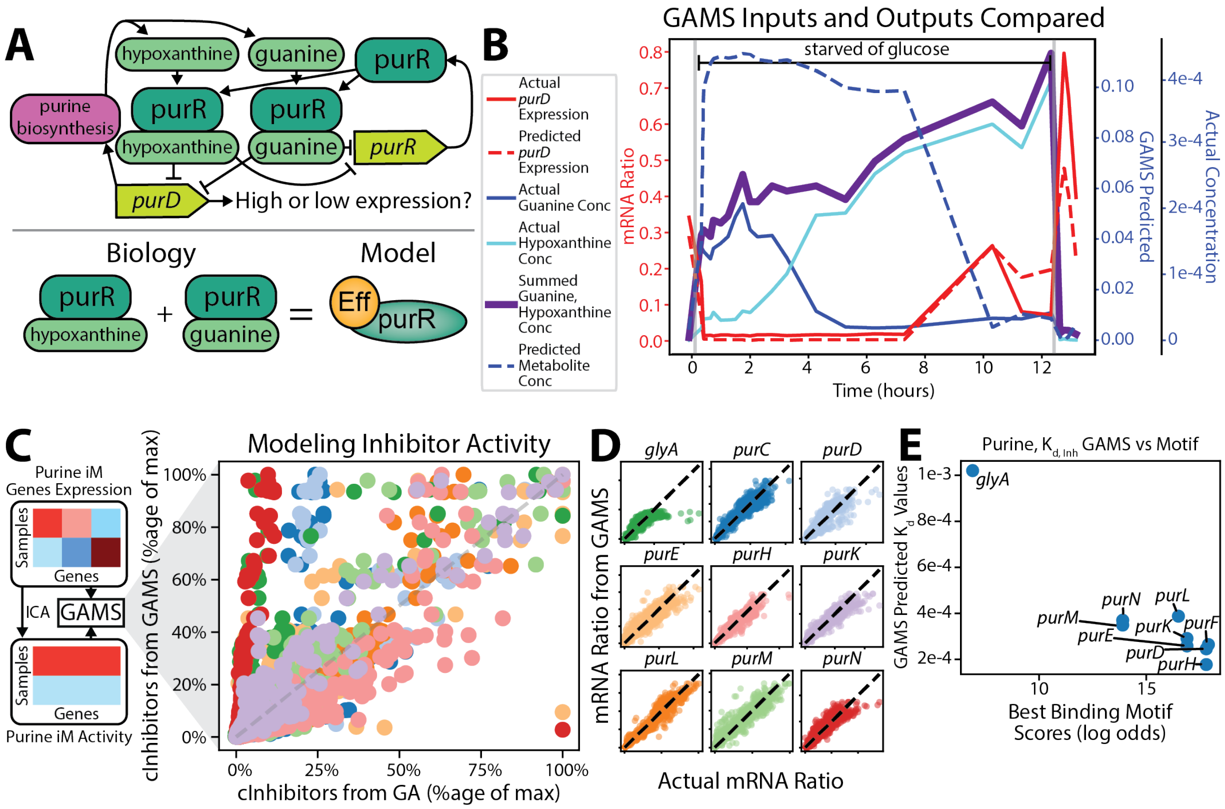

Figure 4.3: **(A)** Regulatory network surrounding *purD* and *purR* involves two separate effector molecules which the products of *purD* and other genes produce. The underlying model treats both effectors as one pooled metabolite concentration. **(B)** Plotted actual and predicted values for metabolite concentrations and gene expression. The model initially accurately predicts both gene expression and metabolite concentration, but fails to accurately model metabolite concentration in the later stationary phase samples. **(C)** GAMS matches the measured and predicted inhibitor activity for the genes in the Purine iModulon. **(D)** The expression of the individual genes modeled in this iModulon can be accurately recreated. **(E)** The sequence-based predictions and model predictions for binding strength of *purR* to the genes it represses are negatively correlated.

4.2.4 *Crp* ALE mutation prediction and validation

We next examined whether the modeling framework could accurately capture transcriptional regulation in cases of promoters affected by multiple transcription factors. We selected genes that appeared in both the Crp-2 iModulon and the Mlc iModulon. *ManXYZ* is activated by Crp-2 and repressed by DhaR, an iModulon with multiple primary regulators, one of which is mlc. *PtsG* is in the DhaR iModulon and is narrowly below the threshold for inclusion in the Crp-2 iModulon, so is included in Crp-2 for this study.

The model for this regulation and effectors involved is outlined in which both regulatory signals are required for accurate prediction of expression, as can be seen in Figure 4.4A. The expansion from one to two regulators requires the additional step of calculating sets of cActivator and cInhibitor that both satisfy the mathematical model and serve as good approximates of regulator activity. As described above, the extension of the workflow to multiple regulators is handled by a first step genetic algorithm which selects a set of valid solutions, which a greedy algorithm can further optimize in order to increase the correlation or anticorrelation between the cActivator or cInhibitor and the iModulon's activity. Figure 4.4B and Figure 4.4C show these steps and how they generate input regulatory activities that satisfy the mathematical model that can be used in the GAMS solver. Figure 4.4DE shows how the GAMS model optimizes to both match the input cActivator and cInhibitor values as well as the expression values.

A *crp* knockout evolution experiment led to convergent mutations upstream of *ptsG* on repressor binding sites (NCBI GEO GSE266148). We tested our model's ability to predict the impacts of these genetic alterations, by taking the control wildtype sample from said *crp* knockout evolution study and setting cActivator to zero to simulate the *crp* KO. Following this, cInhibitor was additionally set to zero to simulate the loss of the repressor. The predicted *crp* KO sample has lower expression than the predicted combined *crp* KO and repressor mutation sample. *Crp* KO strains from this study follow a similar pattern (see Supplemental Figure 4.1).

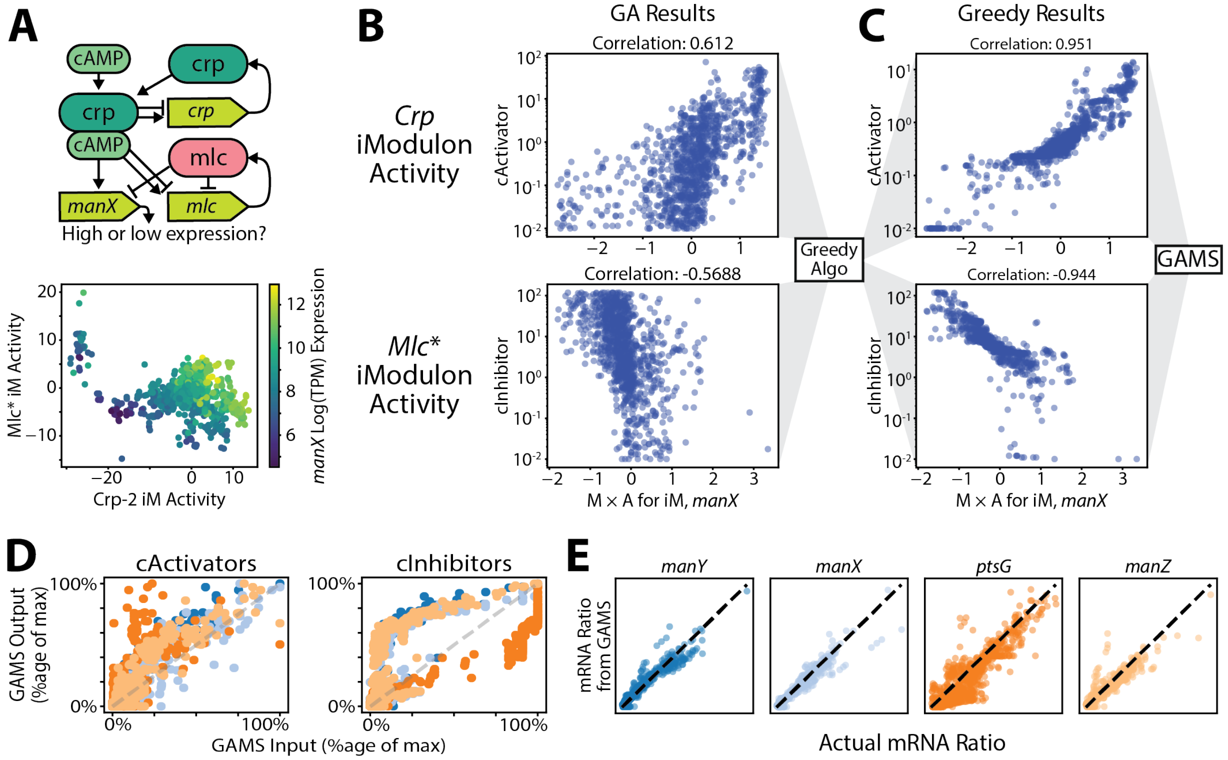

Figure 4.4: **(A)** The surrounding regulatory network of *manX* regulation which is repressed by *mlc* and promoted by *crp*, which also play roles in regulating themselves. The iModulons associated with these regulators determine *manX* expression. The Mlc iModulon is named DhaR in PRECISE1K as both *mlc* and *dhaR* are primary regulators of the same iModulon. (B) Because there is a promoter and inhibitor, there is one more unknown than equations and cActivator and cInhibitor must be calculated using a coupled genetic algorithm (GA). The resulting values are shown. (C) The results of the GA algorithm are then fed into a greedy algorithm which further optimizes the cActivator and cInhibitor values to be fed into GAMS. (D) The GAMS model matches the input cActivator and cInhibitor values from the greedy algorithm. (E) The GAMS model recreates the expression profiles of the modeled genes.

4.2.5 Extending the workflow as a quantitative reconstruction of the *E. coli* TRN

The high overlap between genes in iModulons and regulons enables the established workflow to use ICA-inferred component activities as a proxy for transcription factor regulatory

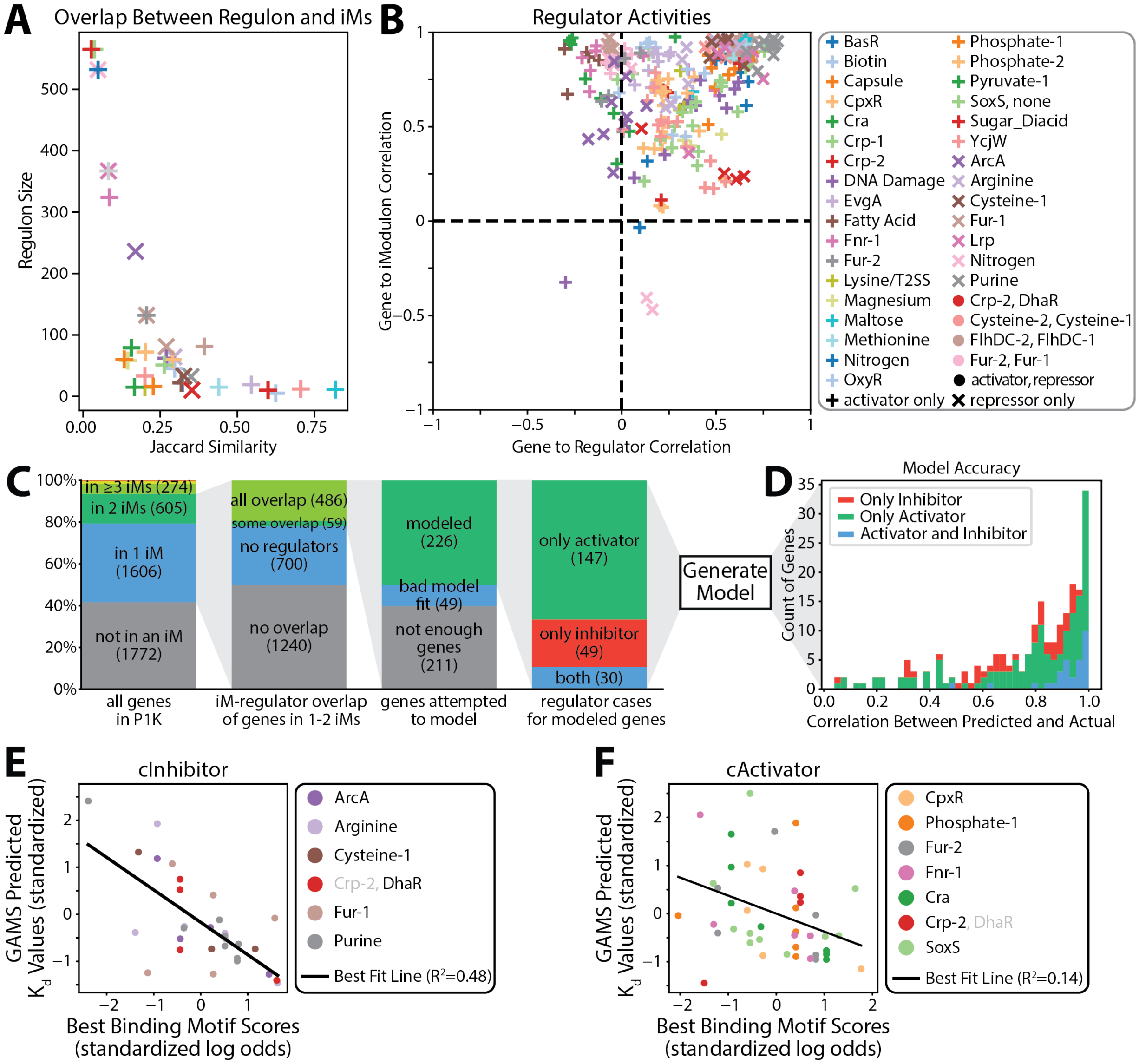

Figure 4.5: **(A)** The amount of overlap between iModulons and the regulons of the regulators of iModulons. High overlap exists for many iModulon-regulator pairs. **(B)** Correlation between genes and their iModulons and regulators. Genes have much higher correlation to their iModulons than regulators. **(C)** The various genes that are unable to be modeled are removed, the reasons for which are outlined. **(D)** The model's accuracy across all samples is about equal for the different regulatory types. **(E)** The GAMS predicted Kd values for the various inhibitors and the sequence-based predictions of motif binding strength are negatively correlated across all samples. Both values are standardized per regulator in order to scale them for comparison. **(F)** The GAMS predicted Kd values for the various activators and the sequence-based predictions of motif binding strength are slightly negatively correlated across all samples. Both values are standardized per regulator in order to scale them for comparison.

activity (Figure 4.5A). Indeed, we observed that iModulons generally have higher correlations to gene expression than does expression of the regulator of the regulons to which a gene belongs (Figure 4.5B). Utilizing this high overlap, we have expanded our workflow to model twelve total regulatory cases. Five are inhibitor only (ArcA, Arginine, Cysteine-1, Fur-1, and Purine iModulons), six are promoter only (CpxR, Phosphate-1, Fur-2, Fnr-1, Cra, Crp-2, and SoxS), and one is dual promoter and inhibitor (Crp-2 and DhaR). In total, these cases account for 226 genes.

Some genes are unable to be modeled for a variety of reasons, most often due to either not belonging to any iModulons or lack of annotated regulator of iModulons to which the gene belongs (Figure 4.5C). For the modeled genes, we are able to accurately recreate gene expression with a median correlation between predicted and actual of 0.82 (Figure 4.5D). The expected negative correlations between sequence-predicted motif scores and GAMS predicted Kd values are found across the modeled regulons, although this component of the model's accuracy is better for repressors than promoters (Figure 4.5EF). Thus, although the scope of the workflow remains somewhat limited compared to the entire *E. coli* TRN, the models that were able to be developed consistently perform well at capturing both gene expression and binding site strength.

4.3 Discussion

Here, we developed a workflow to utilize ICA component (iModulon) activities extracted from gene expression databases as a proxy for regulator activities to parameterize transcriptional regulatory models in *E. coli.* Transcription regulation was modeled with standard thermodynamic binding equations based on well-established promoter complexes incorporating known transcription factor effectors. This workflow generates a data-backed physical model for gene regulation and enables the prediction of metabolites and transcription factor binding constants using expression data. These predictions can be verified using external data sources, thus showing the robustness of this workflow.

The Arginine iModulon, regulated by *argR*, serves as the primary repressor for several genes that control arginine import and biosynthesis. Expression of these genes is positively correlated with expression of their repressor *argR*, a counter-intuitive relationship that the model successfully captures by accounting for the role of the *argR* activator arginine, which governs the effective argR activity level. In another case study, *purR*, which regulates the Purine iModulon, is activated by either guanine or hypoxanthine, [148] and the model aggregates the effect of both effectors explicitly in its formulation. When compared to actual metabolite concentrations, the model is able to predict the sum concentration of both metabolites for the first few hours of a glucose starved dataset [147]. Thus, initial case studies suggest that the modeling formalism developed here can capture complex regulation scenarios. The lack of accuracy for the later hours is possibly due to our model not accounting for sigma factors which have a large effect on stationary phase expression [149]. Much effort has gone into modeling binding strength of sigma factors [150] and proper utilization of this knowledge may enable its inclusion in a mechanistic model such as ours.

Many promoters are regulated by multiple transcription factors and other DNA-binding proteins. We developed a case study to examine the ability of the developed modeling formalism to capture regulator interactions at promoters. A small set of genes are promoted by the Crp-2 iModulon and inhibited by the DhaR iModulon, which are promoted by *crp* and *mlc* respectively. This regulatory case provides an example of our model predicting expression for genes with both an activator and repressor. The model is able to accurately model expression by providing well correlated cActivator and cInhibitor values through the use of genetic and greedy algorithms.

Multi-regulator genes are common, as 1,413 genes have two or more annotated regulator binding sites [137]. To encompass these genes in future versions of this workflow, we will need to expand the number of different promoter states our model accounts for.

This workflow has currently been expanded to 12 regulatory cases which includes 226 genes. 54% of the genes in PRECISE1K account for only 20% of total expression variance [106] and an additional 37% of the remaining genes have no annotated regulator [137]. This in total leaves 1,265 genes that explain significant variance and have known regulators, of which 138 are included in our model. To further expand this model, we will need to include genes not currently in iModulons or include the genes of iModulons that are not directly regulated by known iModulon regulators. Larger-scale models can account for more genes [151], but lack the mechanical details that a workflow such as ours allows. Combining these approaches could potentially yield a detailed mechanistic model for all of transcription.

The mechanistic promoter models presented here connect together detailed biological research about promoter-regulator interaction with the global regulatory trends revealed through large-scale data analytics. This enables the inferrance of biological constants and metabolite measurements which can be directly mapped to gene expression. Whole-cell models exist which are able to connect metabolism to the cell's environment through the utilization of differential equations and numerous parameters [145, 152]. The promoter regulation models developed here could presumably be integrated within multi-scale and whole cell frameworks to better represent transcriptional regulation during life cycle simulations. Taken together, this study suggests a path toward generating a large-scale mechanistic model of gene regulation in *E. coli*.

4.4 Methods

iModulon Computation

RNA-sequencing data was used to create iModulon activity levels of our strains using PyModulon [37] which is available at https://github.com/SBRG/pymodulon. Activities of iModulons were compared to samples from PRECISE1K which is easily accessible using iModulonDB [5]. The calculations of new iModulons for our dataset in addition to PRECISE1K was performed using modulome-workflow [110] which is available at https://github.com/avsastry/modulome-workflow. See https://imodulondb.org/ [5] for a more complete description of iModulons and their calculation.

Workflow Overview

The process described here is outlined in Supplemental Figure C.1. Before any data is processed, genes are selected for modeling (see Methods: Initial Selection of Genes to Model) and outliers samples are removed (see Methods: Removal of Outlier Samples). The initial inputs are gene expression files in log TPM format, the M and A matrices output from ICA, and a proteomic dataset [153]. These three datasets are used to create mRNA expression ratios, recreated gene expression values from M × A, and creating constraints for the model parameters. Independent of this, ideal biological constants, namely $K_{d,RNAP}$, k_escape, and $K_{eq,opening}$ are selected from a calculated set of possible solutions to the mathematical equations (see Methods: Selection of Biological Constants). The M × A values for each gene and these ideal biological constants are used to generate cActivator and cInhibitor values (see Methods: Creating cActivator and cInhibitor Values). If there is both an activating and inhibiting iModulon, a genetic algorithm and following greedy algorithm are used to further improve the cActivator and cInhibitor values (see Methods: GA and Greedy Algorithm Optimization).

These cActivator and/or cInhibitor values are then input into the GAMS optimization software which recreates new cActivator and cInhibitor values calculated from the correct regulator type equation (see Methods: Physical Model Equations). GAMS optimizes two primary criteria, matching to the input cActivator and/or cInhibitor values and matching gene expression values to the actual gene expression values. The GAMS model then outputs the underlying constants for the new cActivator and cInhibitor values as well as the predicted gene expression values (see Methods: GAMS Model). The GAMS model can be rerun multiple times with changing constraints and weightings in order to achieve higher agreement between output and input values (see Methods: GAMS Parameter Optimization).

Initial Selection of Genes to Model

In order for a gene to be included in our transcriptional model, it must satisfy a handful of criteria. First, the gene must be in one or two iModulons that have a well characterized regulator that also regulates the gene. This regulation is determined by the existence of a strong or confirmed regulatory relationship according to RegulonDB [137]. If a gene is in two iModulons, it must be repressed by one and promoted by the other.

At this point, initial biological constants can be generated for the gene and produce cActivator and cInhibitor values. For some genes, there is no valid set of biological constants which result in usable cActivator or cInhibitor values. This is most often due to either extremely high or low expression values for the gene without either violating the underlying mathematical equations or creating negative cActivator or cInhibitor values. The transcription factors themselves, often members of their own iModulons, are also removed. All remaining genes are grouped by their respective iModulons, into their regulatory cases: activator only cases, inhibitor only cases, or dual activator and inhibitor cases After these steps, regulatory cases are removed if they contain only one sample.

Mathematical Model Equations

The basis for the equations are primarily Michaelis-Menten dynamics along with a few transcriptionally related formulas. When solved together this results in:

$$mRNA_{\text{ratio}} = \frac{\left(\text{ cActivator } K_{d,RNAP}+K_{d,Act-RNAP}\right)\left(K_{d,RNAP}+[RNAP]+K_{\text{eq, opening }} K_{d,RNAP}\right)}{(1+\text{ cActivator }+\text{ cInhibitor })K_{d,RNAP}K_{d,Ac-RNAP}+\text{ cActivator } K_{d,RNAP}\left(1+K_{\text{eq, opening }}\right)[RNAP]+K_{d,Aet-RNAP}\left(1+K_{\text{dq, opening }}\right)[RNAP]} \tag{4.1}$$

To incorporate different effector binding types, the formula for cActivator and cInhibitor have a few various forms. The following equation is used for calculating cActivator or cInhibitor in terms of metabolite concentrations and $K_{d,Act}$ in the case that the transcription factor has multiple (3) effector binding sites:

$$\begin{aligned}\text{cActivator } &= \frac{1}{18{K_{d,TF}}^2}(3[\text{ effector }]K_{d,TF}+K_{d,eff}K_{d,TF}+3K_{d,TF}[TF]+\\ &\sqrt{-36[\text{ effector }]K_{d,TF}[TF]+(3[\text{ effector }]K_{d,TF}+K_{d,eff}K_{d,TF}+3K_{d,TF}[TF])^2}\Big)\end{aligned} \tag{4.2}$$

The following equation is used for calculating cActivator in terms of metabolite concentrations and $K_{d,Act}$ in the case that the transcription factor has 2 co-effectors:

$$\begin{aligned}\text{cActivator } &= \frac{1}{4K_{d,TF}}(K_{d,\text{ eff }}+[\text{ effector }]+[TF]+\\ &\sqrt{{K_{d,eff}}^2+([\text{ effector }]-[TF])^2+2K_{d,\text{ eff }}([\text{ effector }]+[TF]))}\end{aligned} \tag{4.3}$$

For the regulatory types involving effectors, TF concentration is set for each individual sample by scaling the overall expression values of the TF to be between the minimum and maximum of a proteomics dataset19.

Removal of Outlier Samples

As the GAMS model uses squared difference regression objectives, outliers can have large effects on the outcome by setting either extremely high or extremely low cActivator or cInhibitor values. To avoid this, a few outlier samples are typically removed for every regulatory case. For each regulatory case, a correlation matrix is created between the samples across the regulatory case's genes. Any sample that does not have greater than or equal to 0.5 correlation value to at least 5% of the other samples is removed. For all regulatory cases, this is always under 4% of total samples.

Selecting Basal Conditions and Conversion to mRNA Ratios

In order to eliminate variables from the equations, mRNA ratios are used as opposed to log TPM expression values. In order to calculate these values, a basal condition must be selected. This condition must be uniform across each regulatory case. For each regulatory case, the expression profiles for the genes of said case are collected and standardized, the resulting distributions are then used to select the basal condition. If the regulatory case is only an inhibitor, the sample with the highest average standardized expression is chosen. If the regulatory case is only an activator, the sample with the lowest average standardized expression is chosen. In a dual promoter and inhibitor case, the sample with the closest to zero average standardized expression is chosen.

All expression values are then un-logged and divided by the un-logged value for their respective basal condition. This generates an mRNA ratio value. The same un-logging and dividing by the basal condition is performed on the ICA reconstructed expression values (M × A) in order to scale them properly so that they can be used to calculate the cInhibitor or cActivator values in the non dual-regulator cases.

Selection of Biological Constants

Some relevant equations need to be defined for this section:

$$[mRNA] = \frac{K_{\text{eq, opening}}\, k_{\text{escape}}\, [\text{ Promoter }]}{\left(\frac{K_{\text{d, RNAP}}}{[RNAP]} + K_{\text{eq, opening}} + 1\right)(u + k_{deg})} \tag{4.4}$$

$$[mRNA] = mRNA_{TPM} \frac{mRNA_{\text{total}}}{10^6} \frac{1}{Volume_{cell}} \frac{1}{N_A} \tag{4.5}$$

The following constants in the above equations are defined as follows: $mRNA_{total}$ = 1800 molecules [154], $Volume_{cell} = 10^{-15}$ L [155], [Promoter] = 10^{-9} M [156], Growth rate u = 1/3600 [156], k_{deg}=ln(2)/300 [156].

A grid of possible solutions for $K_{d,RNAP}$, k_{escape}, and K_{eq}, opening is created which is later tested to determine a best set of values for the model for each gene. First a range of possible values are assumed for $K_{d,RNAP} = 10^{[-7,-5]}$ and $k_{escape} = 10^{[-3,1]}$. We calculate the minimum viable value for k_{escape} assuming the max $K_{eq,opening} = 100$ and update the range of possible values for $K_{d,RNAP}$ and k_{escape}. We generate nine new pairwise combinations of $K_{d,RNAP}$ and k_{escape} which satisfies the equation. The value of $K_{eq,opening}$ is calculated in order to solve the equation when [mRNA] is known. This process altogether creates 9 valid sets of constants which correctly calculate [mRNA].

$K_{d,Act-RNAP}$ is set through a two step optimization process for any regulator case with an activating iModulon. In the first step, $K_{d,Act-RNAP}$ is set to $K_{d,RNAP}$ which creates negative cActivator values. $K_{d,Act-RNAP}$ is then incrementally reduced until there are no longer any negative cActivator values. This sets the maximum value for $K_{d,Act-RNAP}$. The second step improves the distribution of cActivator values to both be between 0 and 1,000 and more evenly spread by maximizing the 80^{th} percentile value of the cActivator values. This is to prevent the creation of extreme outlier cActivator values which GAMS will then overfit on if not corrected.

The selection of $K_{d,Act-RNAP}$ described above is carried out for each of these 9 sets of valid constants and the set with the lowest $K_{d,RNAP}$ which also is able to create a valid $K_{d,Act-RNAP}$ is selected for the gene. For some genes, there is not a valid $K_{d,Act-RNAP}$ value which satisfies the equation without creating negative cActivator values, in which case the gene is removed from the model.

Creating cActivator and cInhibitor Values

If there is only a cActivator or cInhibitor, the same number of unknown variables and equations exist so the value can be set. Instead of directly inputting the mRNA ratio value, we input ICA reconstructed expression values (M × A) in order to better model regulatory activity as opposed to expression. These cActivator or cInhibitor values are then calculated for each sample in the gene and each gene in the regulatory case.

If there is a cActivator and cInhibitor, these values need to be calculated simultaneously. First a set of valid solutions is found with various values of cInhibitor and cActivator, initially ranging between 0 and 1,000. These valid solutions are then passed to the genetic algorithm to select sets that correlate with the inhibiting and activating iModulons.

Genetic and Greedy Algorithm Optimization

If there is both a cActivator and cInhibitor in a regulatory case, the underlying mathematical equations must be solved simultaneously for both cActivator and cInhibitor. To pick valid values of cActivator and cInhibitor that also correlate with their respective iModulon activities and thus regulatory activity, two algorithms are used.

First, a Genetic Algorithm (GA) gives us an initial solution [157] using a modified version of eaMuPlusLambda. The GA used for this study is multi-objective as it attempts to both maximize the Spearman correlation between the cActivator and the activating iModulon's activity and minimize the Spearman correlation between cInhibitor and the inhibiting iModulon's activity (as we want a highly negative correlation for cInhibitor and the inhibitor iModulon). Each individual in the GA consists of a randomly-sampled valid solution for cActivator and cInhibitor for every condition. A population of these individuals is created by random individual creation 100 times. Each individual's fitness is evaluated using the two objectives and ranked based on their performance. The best individuals from each generation are selected to propagate to the next generation using the SPEA2 algorithm [158]. 5% of the selected individuals undergo mutation in which a random cActivator-cInhibitor pair is selected for a random amount of conditions in the individual. An additional 5% of the selected individuals undergo crossover in which a random condition's cActivator-cInhibitor pair is swapped between two individuals.

The values generated by this algorithm do drastically improve the positive and negative correlations for cActivator and cInhibitor, but an additional greedy algorithm further optimizes the solution. This algorithm searches the local cActivator-cInhibitor pairs to find better ones using local gradient information. The order of the conditions are randomly shuffled for the best individual in the current generation. Each other condition is iterated over to see if there is a better scored cActivator-cInhibitor pair within 10 steps. This process is then carried out for each condition. The entire greedy algorithm is repeated 50 times. This results in highly correlated and negatively correlated cActivator and cInhibitor values that are also valid solutions to the underlying mathematical equations.

Selection of Constraints on the GAMS Model

The variables that GAMS uses to calculate cActivator and cInhibitor (and thus gene expression) are initially constrained by experimentally derived data. For $K_{d,Act}$ and $K_{d,Inh}$ this constraint is initially based experimentally derived constants for crp [159]. The initial constraints are the minimum observed Kd value for crp divided by 1,000 and the maximum observed Kd value for crp multiplied by 1,000 times. The metabolite concentrations are constrained by the minimum observed concentration divided by 1,000 and the maximum observed concentration multiplied by 1,000 [147]. If no such measurement of the metabolite exists in Link et al. 2015, the minimum of any metabolite in said study divided by 1,000 is used as the minimum and the maximum of any metabolite in said study multiplied by 1,000 is used as the maximum.

GAMS Model

The GAMS part of the workflow is a multi-parameter optimization model that optimizes two objectives: 1) matching actual and predicted mRNA ratios, and 2) matching actual and predicted regulator activities. The first step is reading in of constants and constraints for modeled parameters, which are previously generated as well as a weighting factor to balance between the two objectives. The input cActivator and cInhibitor values are also input, along with a mapping of what type of regulator each iModulon is modeled as. The model variables are created, constrained, and populated with initial values.

The dnlp solver is used for our model. The optimization is then run to minimize the squared error of both objective functions. Once complete, the following modeled variables are output: predicted mRNA ratio, cActivator, cInhibitor, and the various variables that cActivator and cInhibitor are calculated based on. Depending on the regulator type, these variables are some combination of effector concentrations, regulator concentrations, $K_{d,Act}$, and $K_{d,Inh}$.

GAMS Parameter Optimization

A few important constrained variables and the weighting between the two criteria have a large impact on model performance. These constrained variables, depending on the regulator type, are effector concentrations, regulator concentrations, $K_{d,Act}$, and $K_{d,Inh}$. In order to pick constraints that lead to more accurate models while still maintaining biologically-relevant constraints, we have developed a parameter optimization pipeline.

This pipeline creates a set of values for each of these important inputs to GAMS. This set contains one value that is equal to the default value, another that is X% higher, and a third that is X% lower with X starting at 200. GAMS is then run on each combination of these sets (so N^3 runs where N is how many parameters are tested). The resulting run with the least squared error for the objective functions is chosen. If there is a tie, usually due to loose constraints, the run with the tightest constraints is chosen. The process above is repeated with the values from the chosen run now being the default values and X being scaled down by 20% of its previous value. This process is repeated until no more large reductions in the error are found.

Motif Analysis

The motif analysis was performed using https://github.com/SBRG/bitome [139]. The motif_search function was used for each using regulonDB as a source for finding the ideal binding motif to scan with. For each transcription factor, each of its regulated genes were scanned from 100 bp upstream to 50 bp downstream for a matching binding motif.

Acknowledgements

This work was supported by the Novo Nordisk Foundation (NNF20CC0035580). We would like to thank Archana Balasubramanian for helpful discussions on this manuscript.

Chapter 4 in part is a reprint of material submitted to *bioRxiv*:

- **C Dalldorf**, G Hughes, G Li, BO Palsson, DC Zielinski. 2024. “Data-driven modeling of bacterial transcriptional regulation” The dissertation author was the primary author.

Chapter 5

Conclusions

Large scale data analysis techniques are now vital to exploiting the increased scale of publicly available data and knowledge. RNA-sequencing data can inform us of how cells adapt their gene expression to their environments which can both inform better strain engineering design principles as well as drastically improve our understanding of infectious diseases. An improved understanding of how genes regulate their genes can be achieved through the use of non-biased data analysis techniques and validated through the comparison to verified biological knowledge.

The introduction describes the increase in the volume, availability, and centralization of publicly available biological data and knowledge, specifically RNA-sequencing data. This is largely due to a reduction in cost per sample of generating said data. The development and application of novel data analysis techniques to utilize these new large central sources are paramount for pushing systems biology forward.

In Chapter 2, we discuss a study investigating the common RNAP mutations found across laboratory evolutions and their impact on the transcriptome. Independent component analysis is able to find that these mutations nearly universally downregulate stress-response genes regulated by RpoS while upregulating ribosomal subunits. Some mutations had location-specific effects that were adaptations to the specific environments from which they were originally found. This tradeoff appears to be conserved across numerous bacterial strains.

Chapter 3 explores a transcription factor knockout and evolution study carried out on glucose minimal media. The unevolved and evolved samples were sequenced, expression profiled, and phenotyped for substrate readiness. This study found multiple adaptation strategies exist for adjusting to the loss of a regulator, largely dependent on the activity of the transcription factor on glucose minimal media and the size of its regulon. For most transcription factors, growth was recovered without mutations specific to the removed regulator. For the *crp* and *fur* knockout strains, growth recovery involved convergent mutations to elements of their own regulatory networks that restore the expression of highly growth important genes while leaving the majority of the removed regulator's regulon abnormally expressed. This shows the robustness of the transcriptional regulatory network of *E. coli* and its ability to quickly adapt to most large perturbations.

Chapter 4 provides a framework and large forward step in the development of biophysical models for gene regulation. The model created through this work is able to accurately predict biological constants such as binding strengths and metabolite concentrations by utilizing data-driven sources of regulatory activity. This work provides quantitative insights into TRN function and is able resolve apparent paradoxes within regulatory networks.

Overall, these projects have helped advance our understanding of transcriptional regulation in bacteria by exploring different adaptation mechanisms of said network and additionally tests this knowledge by modeling its behavior. Future steps are likely to involve expansion of the regulatory model and a more clear understanding of how independent component analysis successfully decomposes regulatory signals. These additional steps would help further connect the gap between actionable biological understanding and the findings of large-scale data analytics. This thesis aims to connect together different knowledge sources using data analytics to bring the scientific community a step closer to a complete understanding of bacterial gene regulation.

Appendix A

The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions - Supplementary Information

A.1 Other RNAP mutations of special interest from our study

Some mutations predicted to have a specific stress response show no clear existence of one in our study. *RpoC* H419P was commonly found in octanoic acid tolerance studies and thus RNA-sequencing data was gathered for it both on M9 and on octanoic acid. Initially the *rpoC* H419P mutants had trouble growing on octanoic acid until the concentration was lowered, so perhaps an octanoic acid-specific adjustment would have been seen if a higher acid concentration was used. *RpoC* H419P has been previously introduced into *E. coli* and shown to be an adaptation to octanoic acid\cite{Chen2020-rg}, so this issue is likely limited to our study.

RpoB I966S was commonly selected for in heat tolerance studies\cite{Gonzalez-Gonzalez2017-uz,Rodriguez-Verdugo2016-fd,Tenaillon2012-wv} and thus we chose to grow it in our study on high temperature conditions. These heat tolerance studies, however, were all carried out using *E. coli B-strains* and there was little effect of this mutation on the RNA-sequencing data so it appears this is a strain specific adaptation.

A.2 Laboratory evolution creates condition-specific convergent mutations

The evolutionary pull towards growth, and thus greed, in ALEs necessitates condition-specific convergent mutations (**Supplemental Table A.9**). These mutations are not strictly limited to RNAP. For example, *oxyR* is a common mutation target for evolution in oxidative stress\cite{Anand2020-wy} and a *topA* mutation was a convergent target in a heat tolerance evolution\cite{Tenaillon2012-wv}. This trend is widespread among ALEs, as 89% of all evolution experiments in ALEdb contain at least one gene that is mutated in 50% or more of their endpoint strains (calculated by looking at the mutations in all evolved *E. coli* experiments in ALEdb). If excluding RNAP genes, this drops to 80%. While RNAP mutations have the ability to favor growth across a wide variety of conditions, different genes are often better mutational targets for specific conditions.

A.3 The fear vs. greed tradeoff is found in WT and across growth conditions

Fear vs. greed changes are not limited to mutations acquired during evolution, as **Supplemental Figure A.11** shows how the transcriptome composition falls on the tradeoff line in nutrient limited growth changes. Furthermore, when limiting nutrients drive the culture into the stationary phase, a time series of points shows how the transcriptome composition moves down the f/g tradeoff line. This movement shows how lower growth rates on entry into the stationary phase comes with both a drastic change in transcriptome composition and an increase in stress readiness as has been shown before\cite{Houser2015-ds}. However, stationary phase cells often have a disconnect between their transcriptomic and proteomic compositions, so the small iModulon movements between the stationary phase samples should be viewed with some skepticism\cite{Houser2015-ds}. The f/g tradeoff is thus reflected in the various growth states of the WT strain as well as a transition in physiological states.

A.4 Supplemental Figures

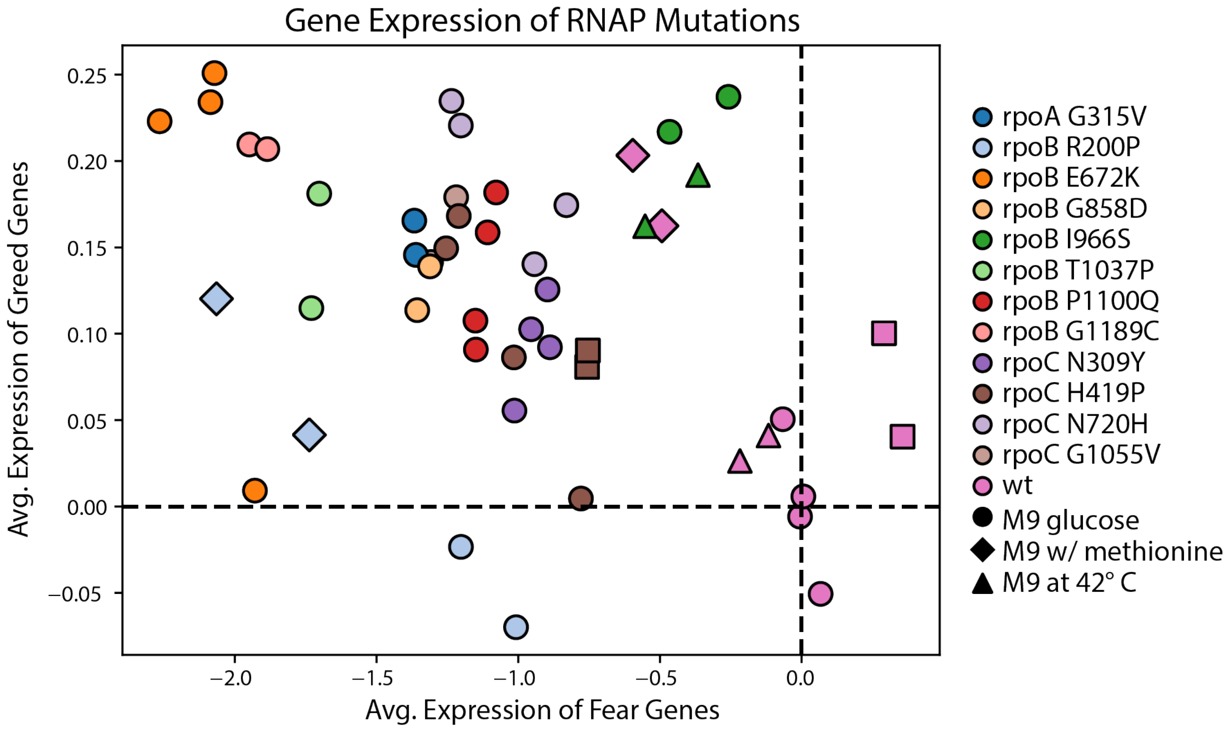

Figure A.1: The average expression values of all replicates from this study are shown for "fear" and "greed" genes. Gene expression is in log TPM and centered on the "wt M9 glucose" condition. "Fear" genes are those that are included in the RpoS and GadX iModulons while "greed" genes are those in the Translation and ppGpp iModulons.

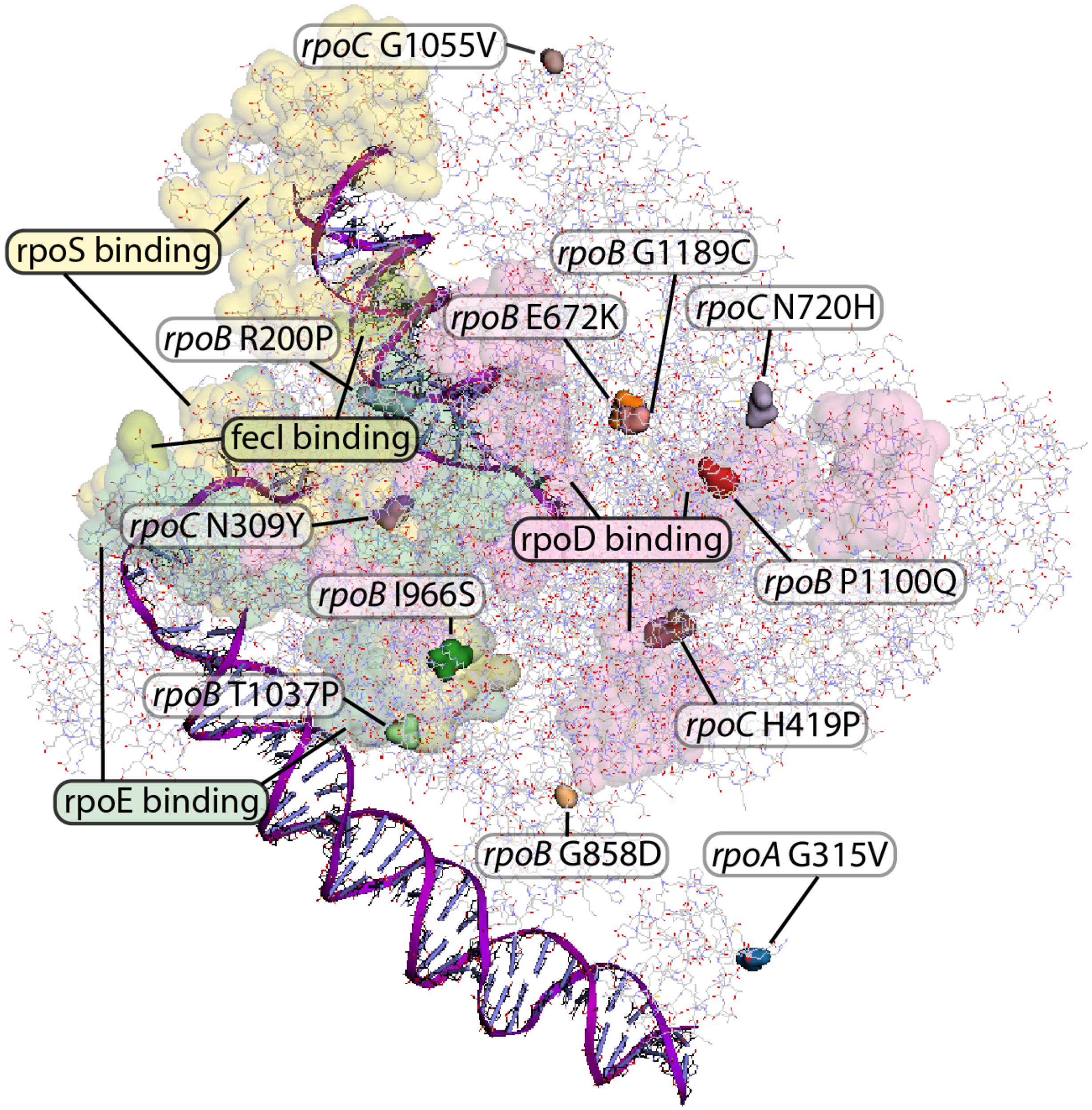

Figure A.2: The structure of RNAP (PDB 6OUL\cite{Chen2019-mm}) is visualized using PyRosetta\cite{Chaudhury2010-uo}, showing the location of mutations used in this study and highlighting some specific RNAP regions of interest. Binding sites are inferred based on the highlighted RNAP residues being within 5 angstrom of said sigma factors in the structural files for RNAP binding with fecI (6JBQ), rpoD (6PST, 6PSR, 6PSQ, 6XLl, 6XL5, 6XL9), rpoE (6JBQ), and rpoS (5IPL, 6KJ6, 6OMF).

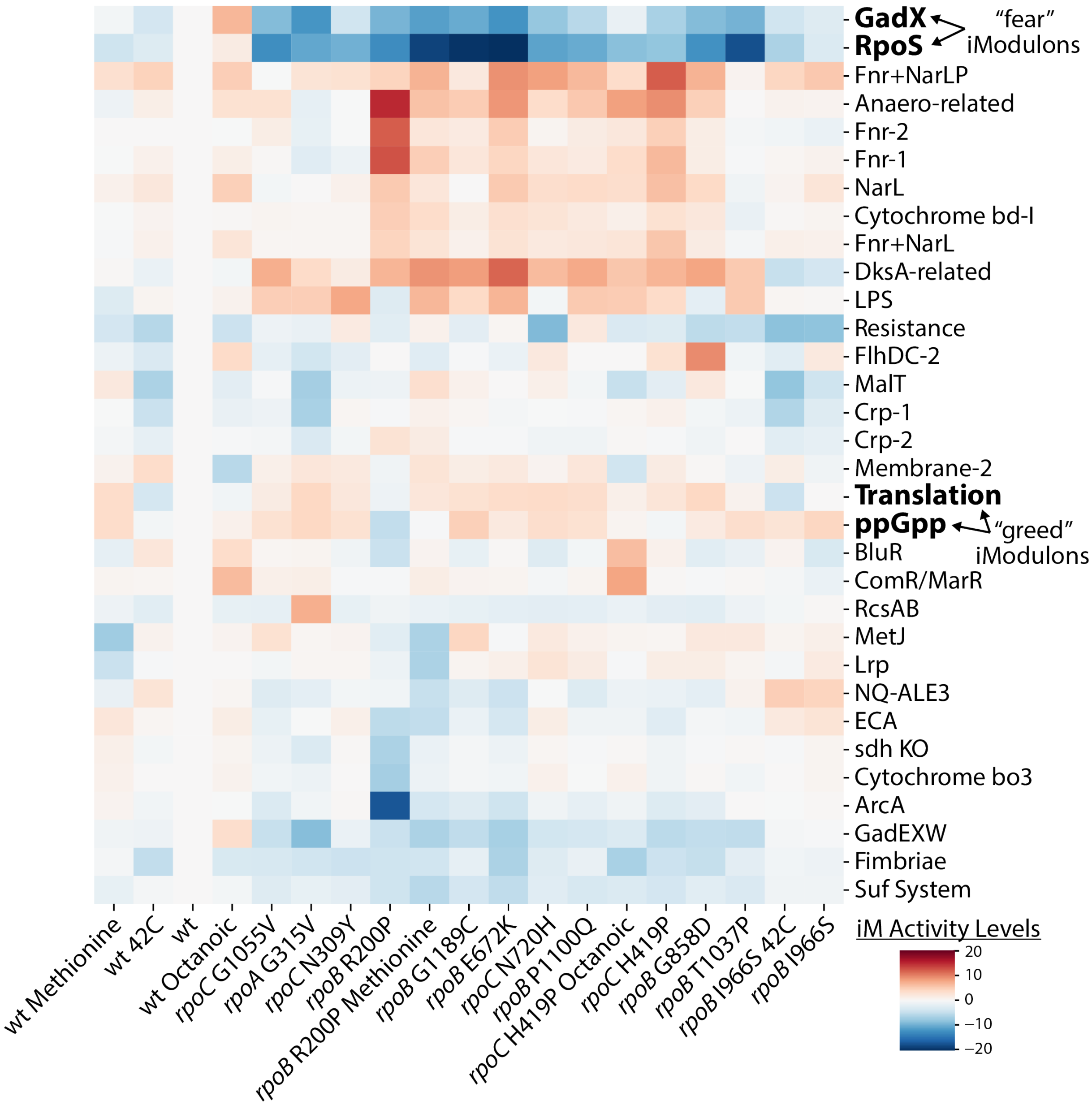

Figure A.3: The most differentially activated iModulons for the mutations and reference conditions are shown here. While RpoS is the strongest effect, some other iModulons are modified. The 30 highest variance iModulons for the listed samples are shown along with Crp-1 and Crp-2.

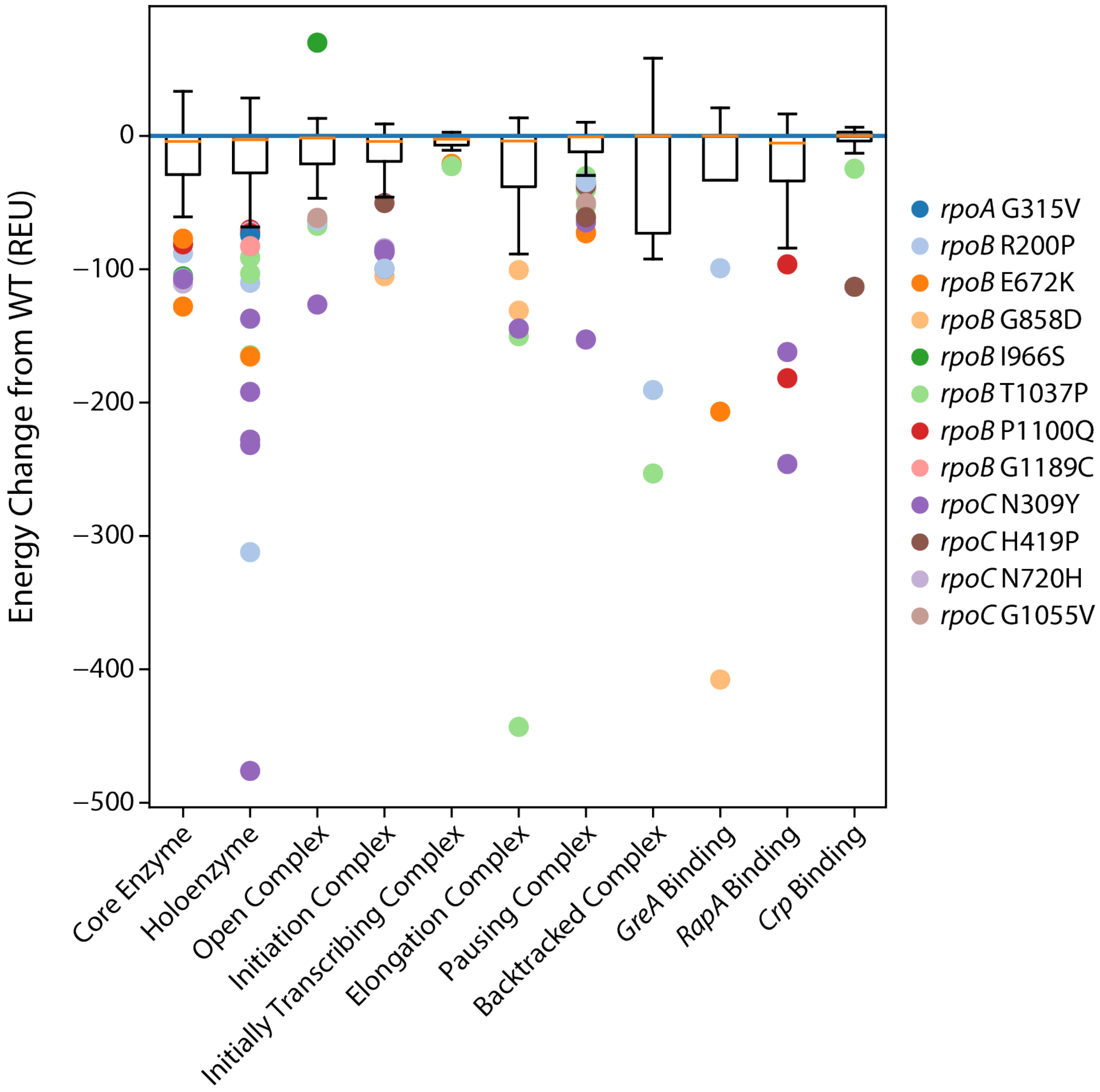

Figure A.4: Nearly all mutations destabilize all of the complexes, with some mutations preferentially destabilizing certain forms. Outliers from boxplot are shown with the box representing the middle two quartiles and the whiskers stretching to 1.5 times the interquartile range.

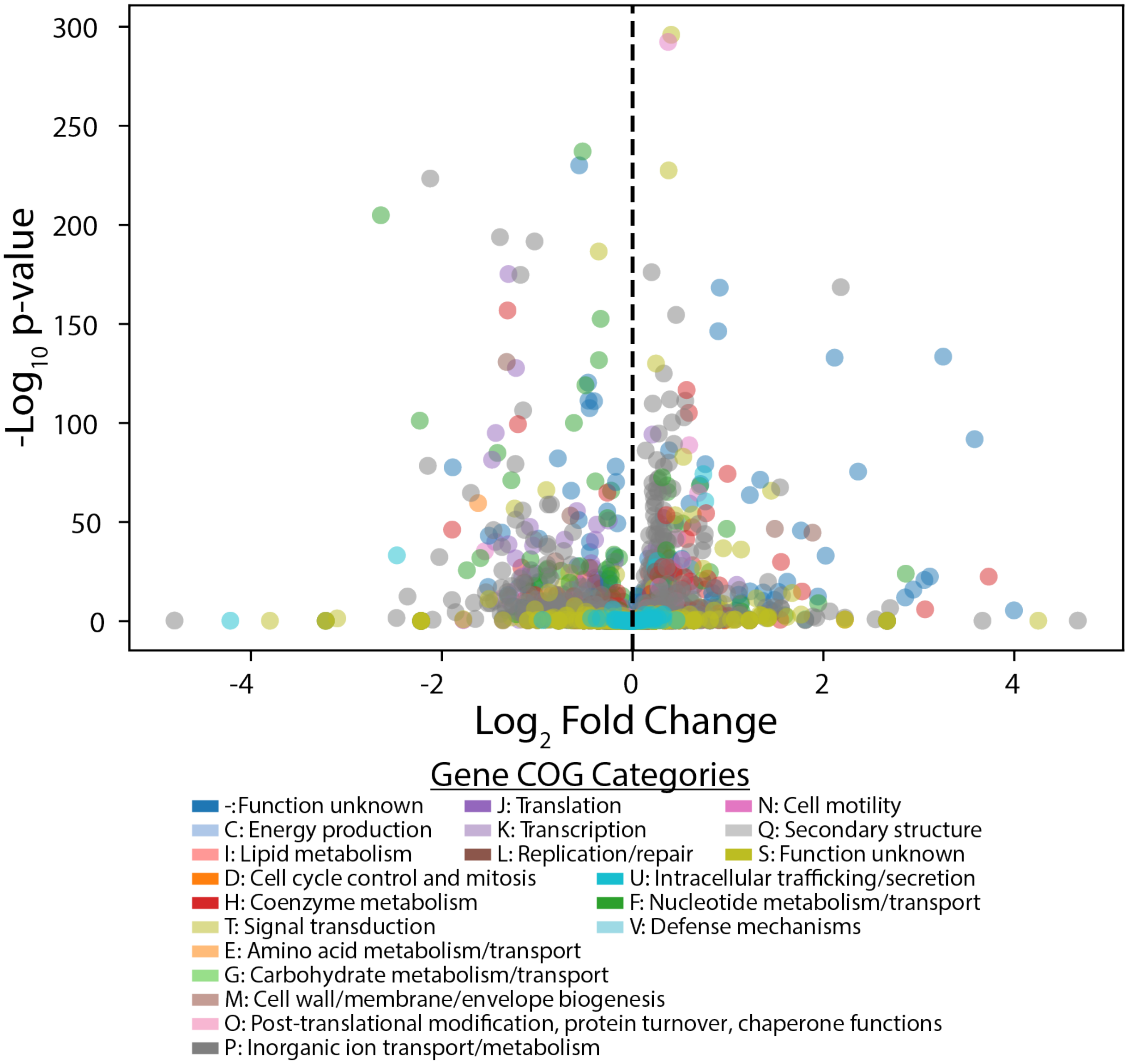

Figure A.5: The median expression value from the mutated strains was used for the mutated strain values. The pairwise single mutant strain compared to the wild-type strain versions of this plot look similar. Interpreting these individual plots is highly difficult and doing so for all these plots together is nearly impossible, thus necessitating the use of iModulon analysis.

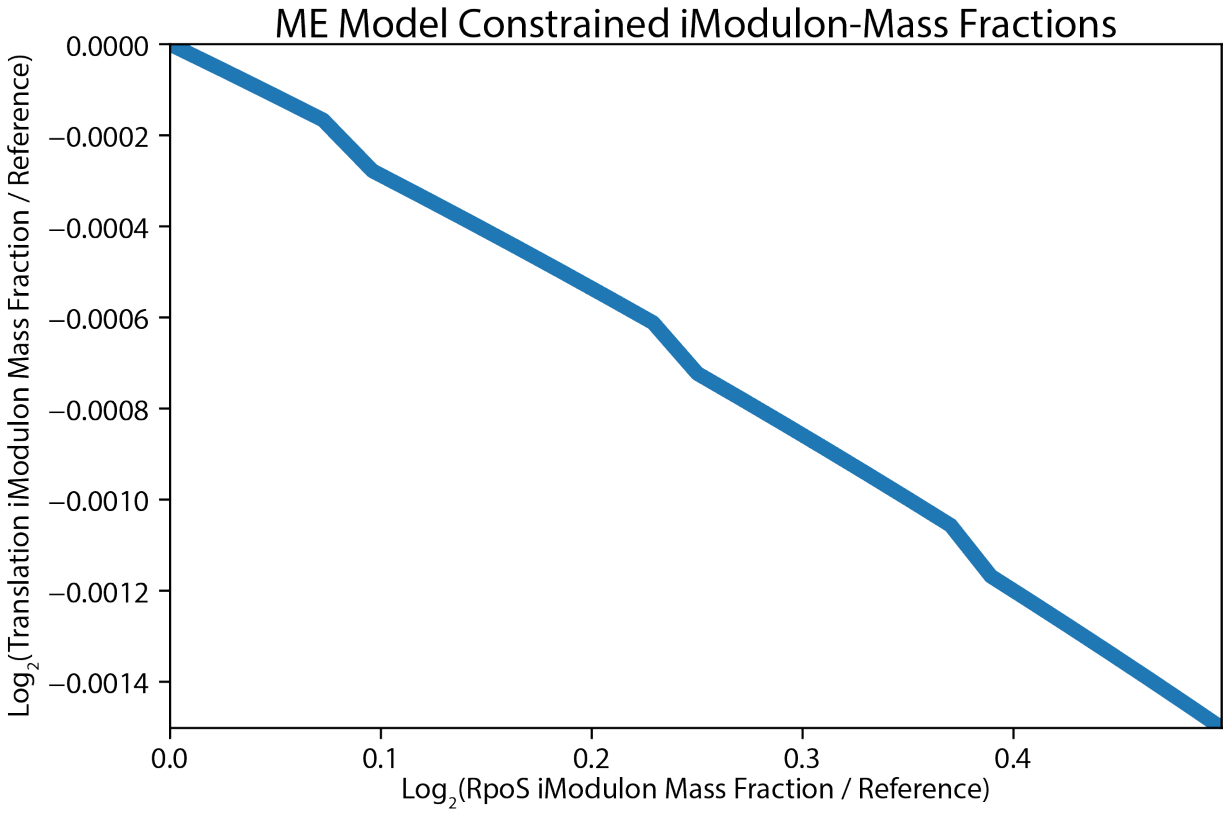

Figure A.6: Reactions associated with the Translation iModulon's genes were tightly controlled in a ME model simulation, resulting in a corresponding change in the proteomic mass fraction in both the Translation iModulon and the RpoS iModulon. Growth rates increased nominally (¡1%) as RpoS decreased. Note that given the large proteomic fraction allocated to the Translation iModulon compared to that of the RpoS iModulon, its fold-changes are numerically much smaller, but represent a notable proteome reallocation.

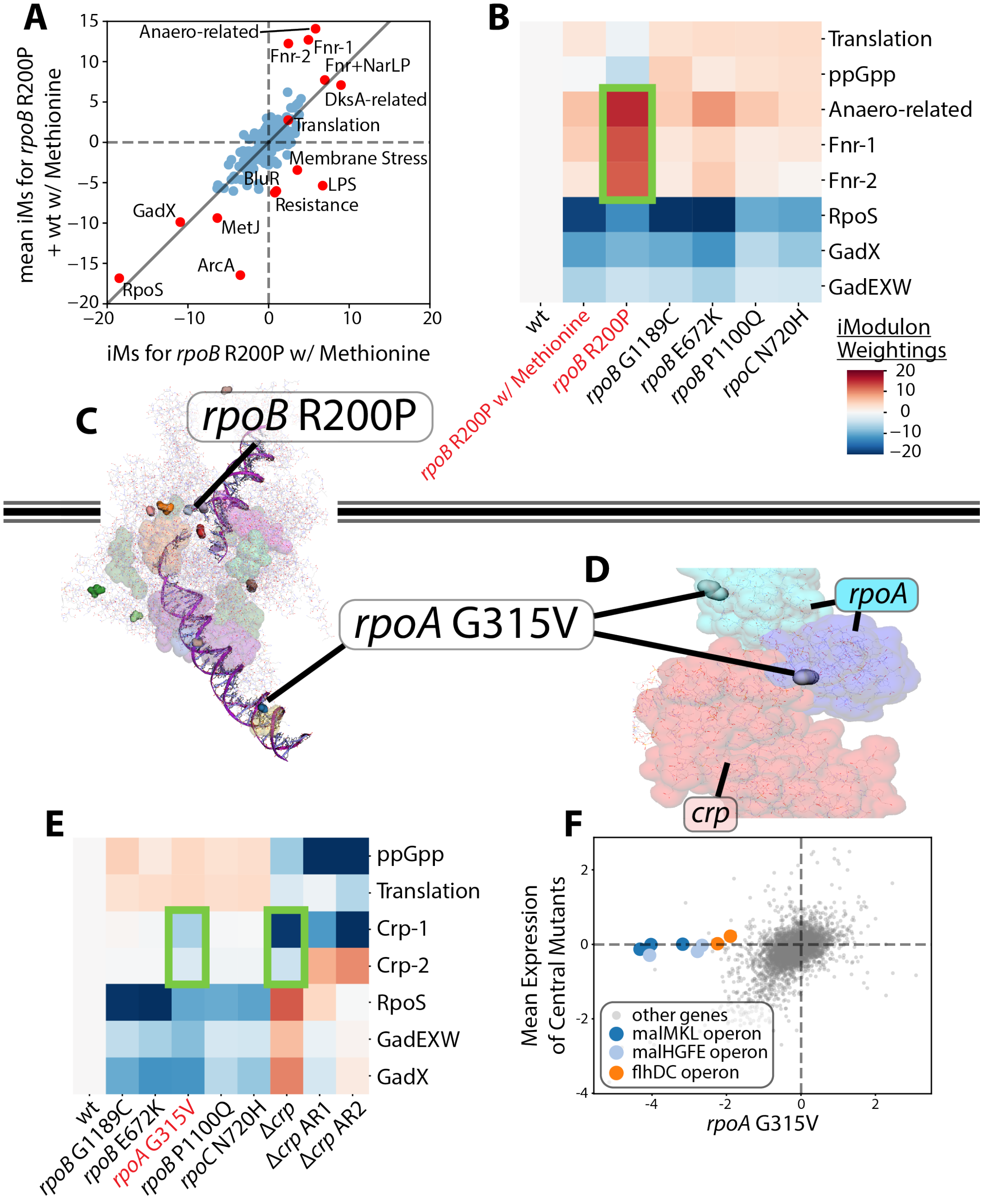

Figure A.7: **(A)** iModulon activities of *rpoB* R200P, wt methionine, and *rpoB* R200P methionine. **(B)** *RpoB* R200P's effect on Fnr and Anaero-related iModulons. **(C)** The location of the two mutations in the protein (PDB 6OUL [61]). **(D)** Location of the *rpoA* mutation in relationship to *crp* (PDB 3N4M [160]). **(E)** *RpoA* G315V's effect on iModulons, compared to *crp* modified strains [71]. (F) *RpoA* G315V compared to the expression of the central mutants.

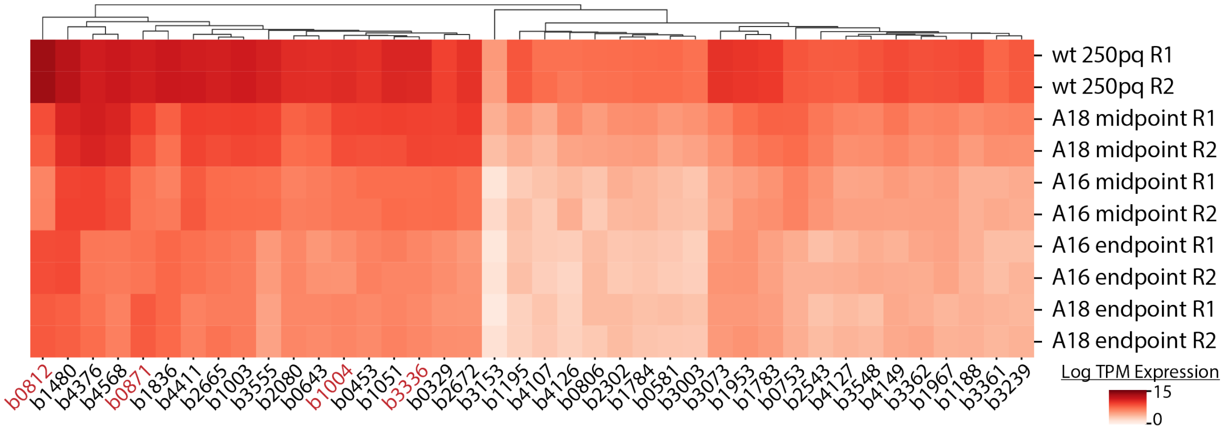

Figure A.8: As the cells evolve on 250 µM paraquat, many genes are downregulated from the RpoS iModulon to enable higher growth, but those related to oxidative stress are not (those highlighted red). Genes listed here are the top 40 most variant within the RpoS iModulon for these strains.

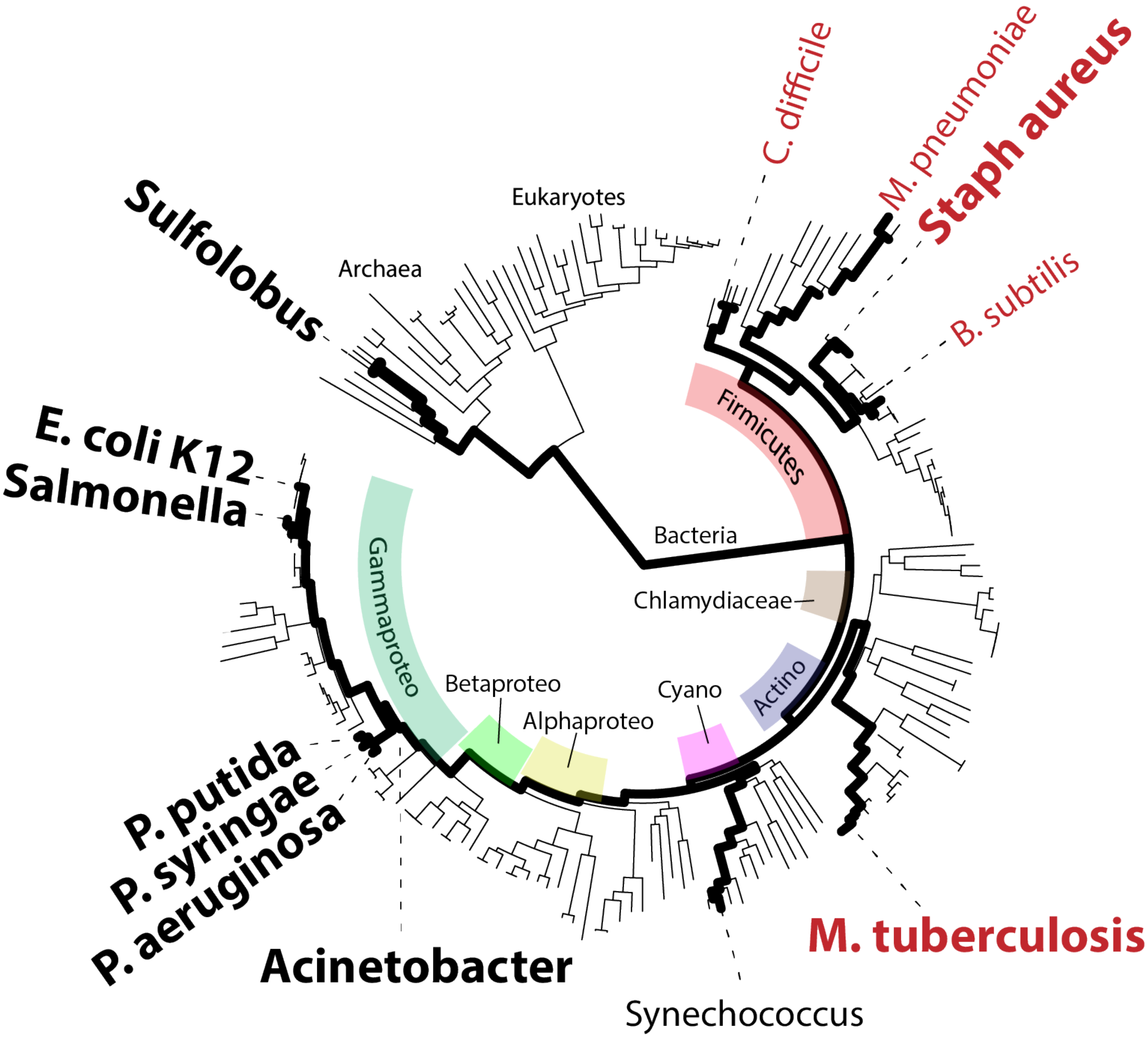

Figure A.9: All species named here were investigated for the existence of said fear vs. greed tradeoff. Red species names are gram positive, black are gram negative. The larger and bolded species names are those that are shown in Fig. 2.5. Species not shown in said figure typically did not have one clear stress iModulon, making a two dimensional visualization of the fear vs. greed comparison difficult.

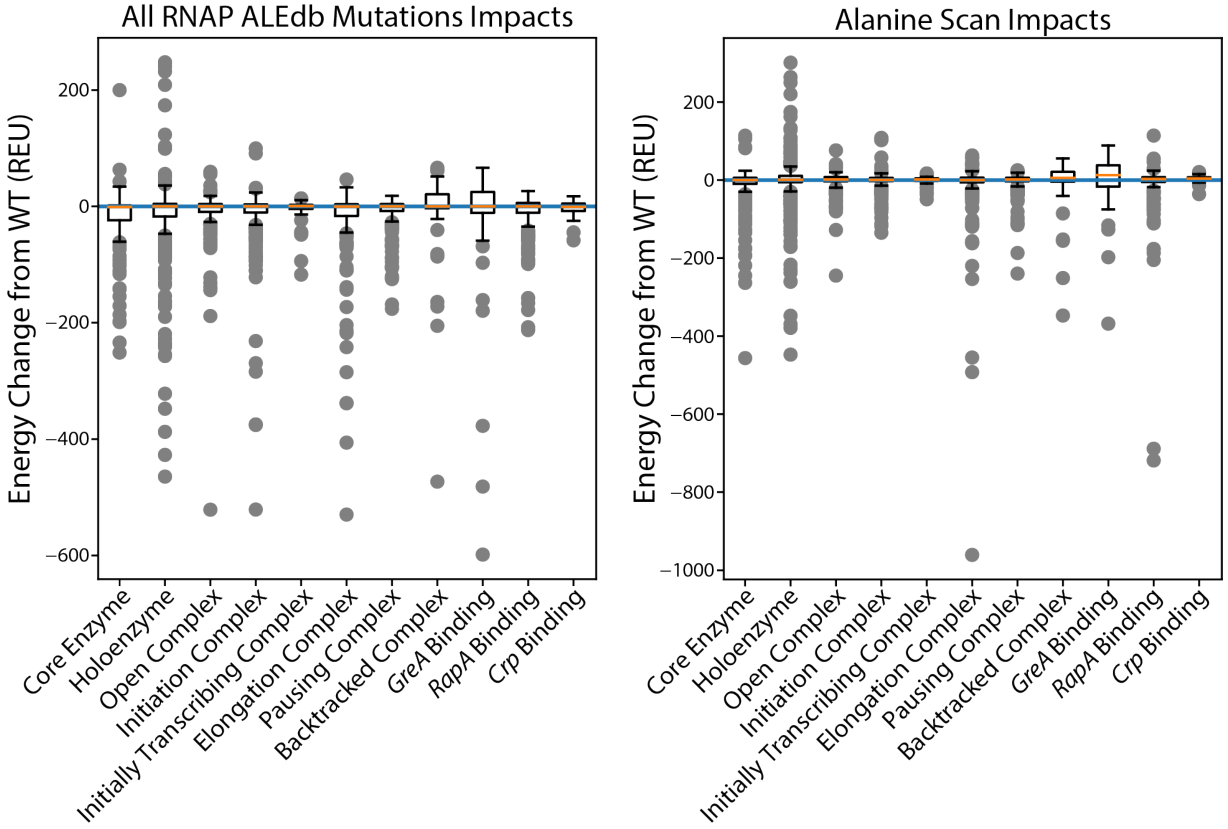

Figure A.10: PDB structures used for each form of RNAP. Outliers from boxplot are shown with the box representing the middle two quartiles and the whiskers stretching to 1.5 times the interquartile range.

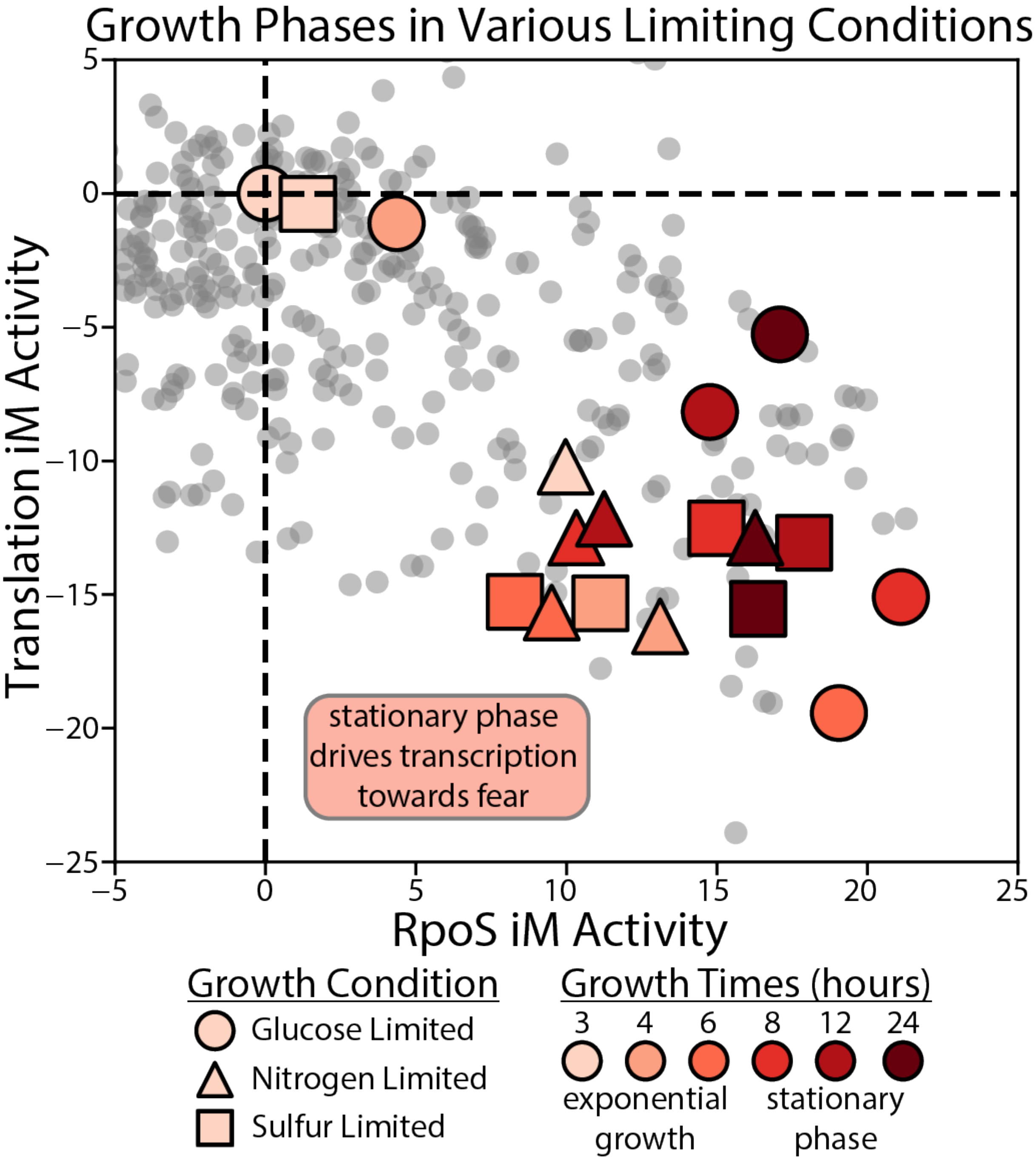

Figure A.11: This transition comes with down regulation of the Translation iModulon and upregulation of stress iModulons. This behavior shows clearly how growth-related genes are downregulated and stress readiness increases in transition to stationary phase, the opposite change to what happens during laboratory evolution to high growth rates. Data is from NCBI GEO GSE226643.

Appendix B

Diversity of transcription regulatory adaptation in ***E. coli*** - Supplementary Information

B.1 Supplemental Figures

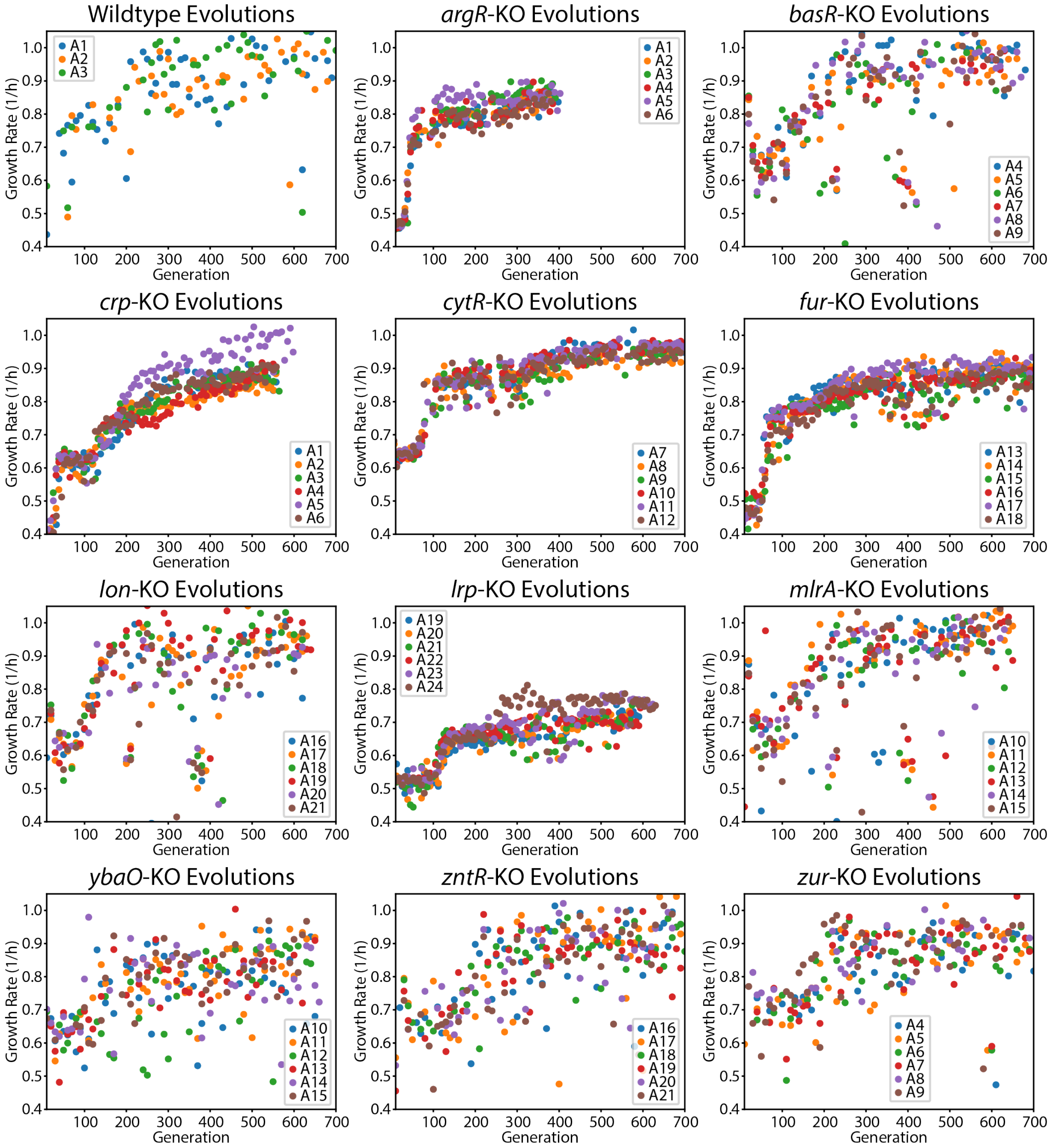

Figure B.1: The independent lineages of each of the 11 TF KO-ALE's and the wildtype samples are shown. Some experiments, such as *argR*, were stopped early due to growth rates approaching wildtype and a lack of growth change.

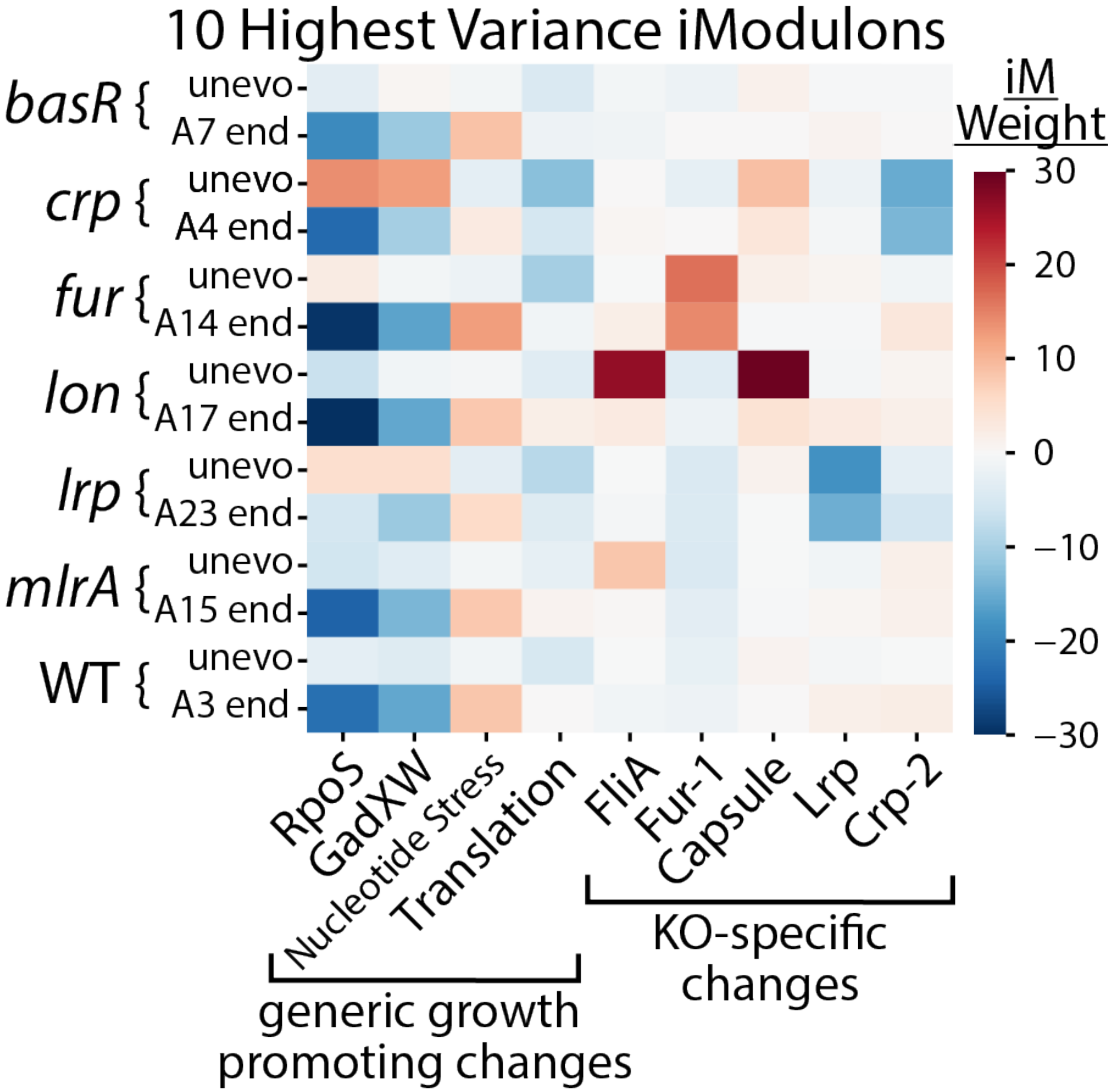

Figure B.2: The top nine iModulons with the highest variance activity among the samples from this study are shown, alongside the Translation iModulon.

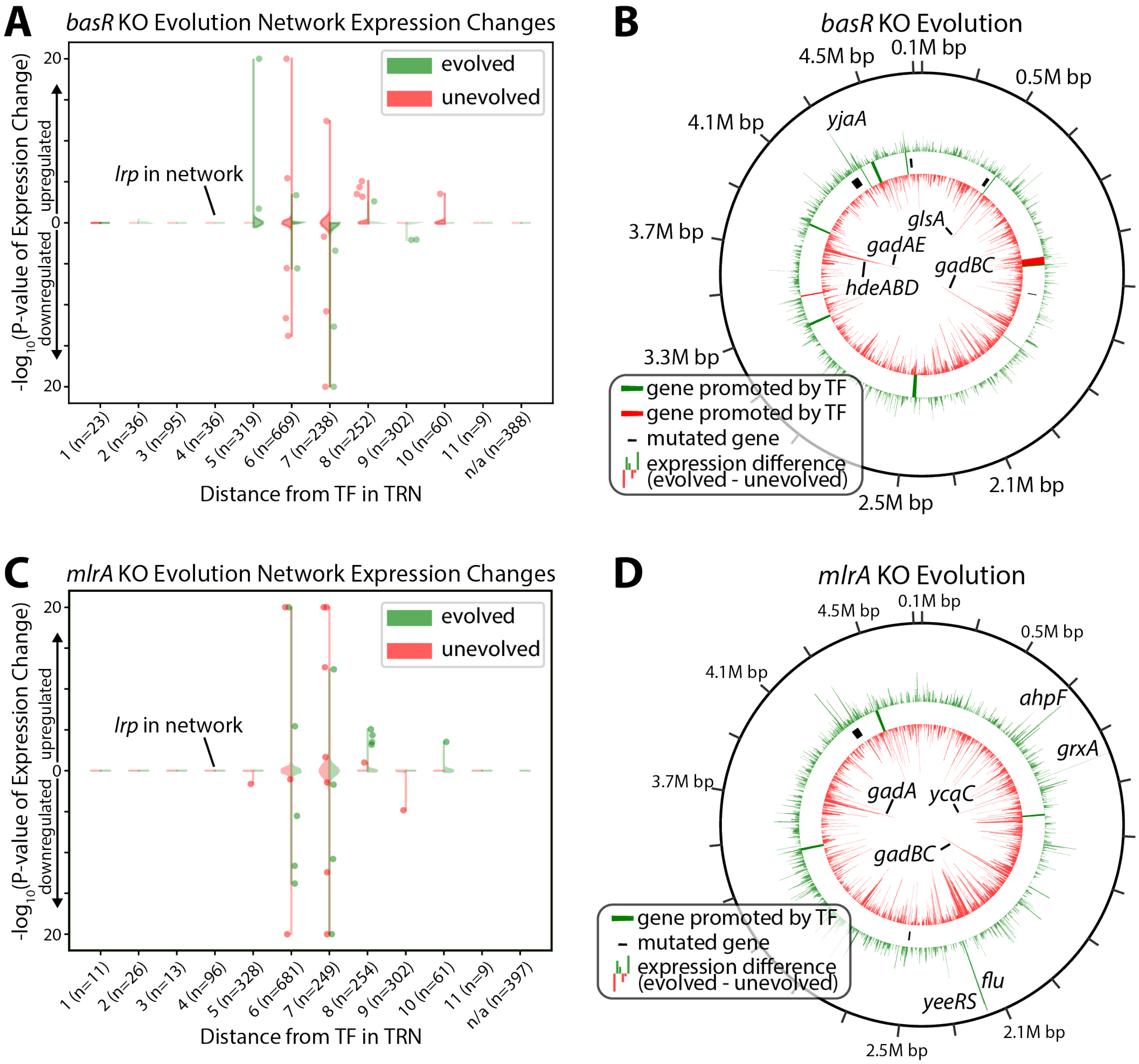

Figure B.3: **(A)** The vertical axis shows the expression change in genes when compared to the unevolved wildtype sample, with the red and green violin plots representing the unevolved and evolved *basR* KO-ALE respectively. Moving from left to right, the first column are all the directly regulated genes of *basR*, the second column are all regulatory targets of the first column's genes, and so on until no more genes can be reached leaving 388 genes that are not connected to *basR*. **(B)** The middle ring shows where in the chromosome mutations and *basR*'s regulon are while the red and green activities represent transcriptional changes for the *basR* KO-ALE. **(C)** Same as panel A, but constructed for *mlrA*. (D) Same as panel B, but constructed for *mlrA*.

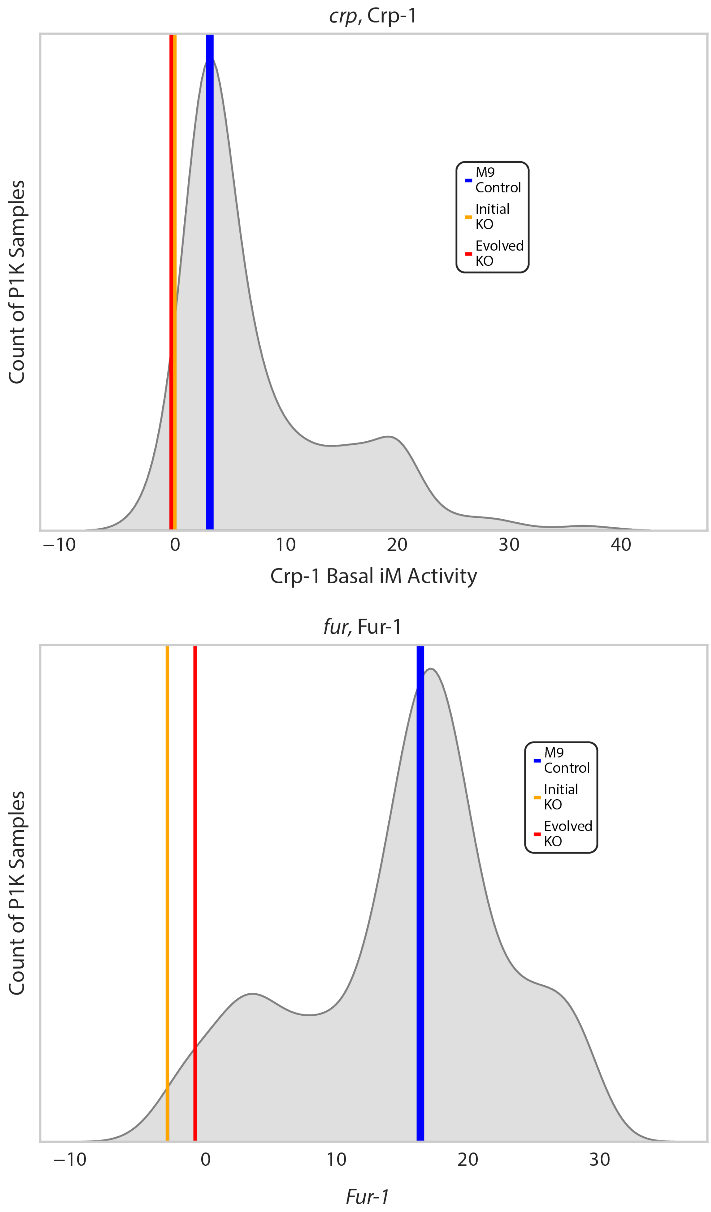

Figure B.4: Density plot of the Crp-1 and Fur-1 basal-adjusted iModulon's activity level across PRECISE1K samples.

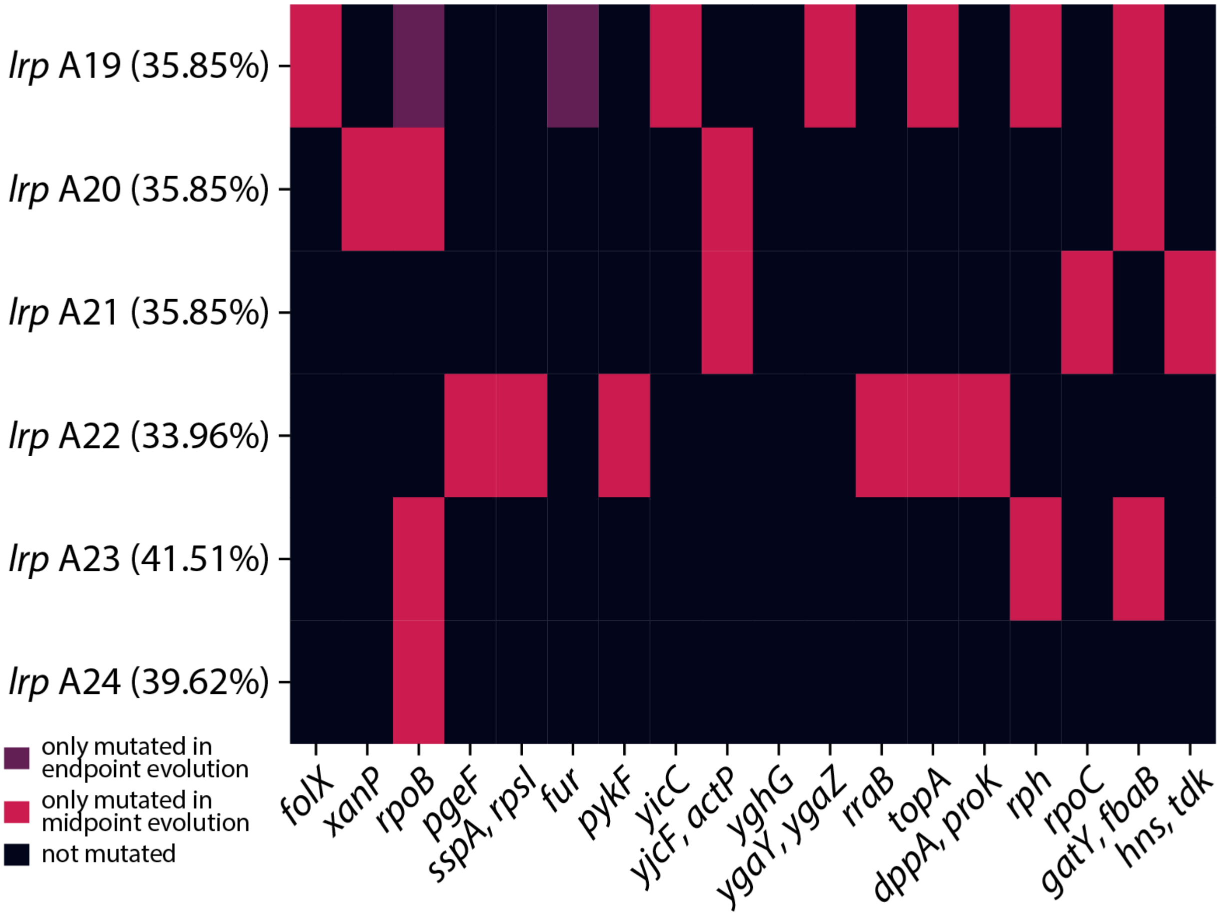

Figure B.5: Mutations found in lrp KO-ALEs across all samples. Only mutations found in at least two samples are shown.

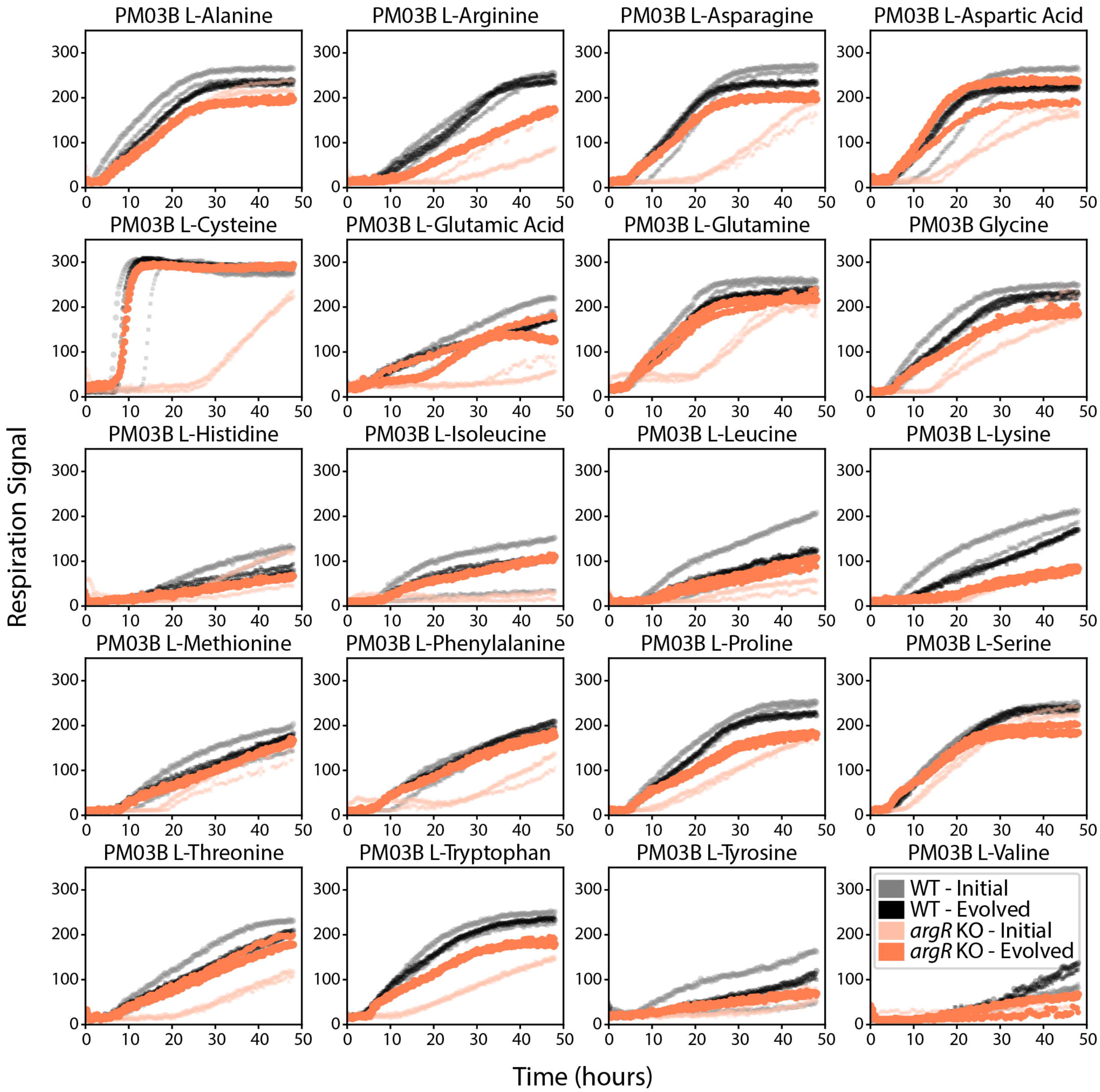

Figure B.6: Respiration signals is an opacity reading that is the measurement for growth for OmniLog plates. Each plot here represents a different nitrogen-limited plate supplemented with an amino acid.

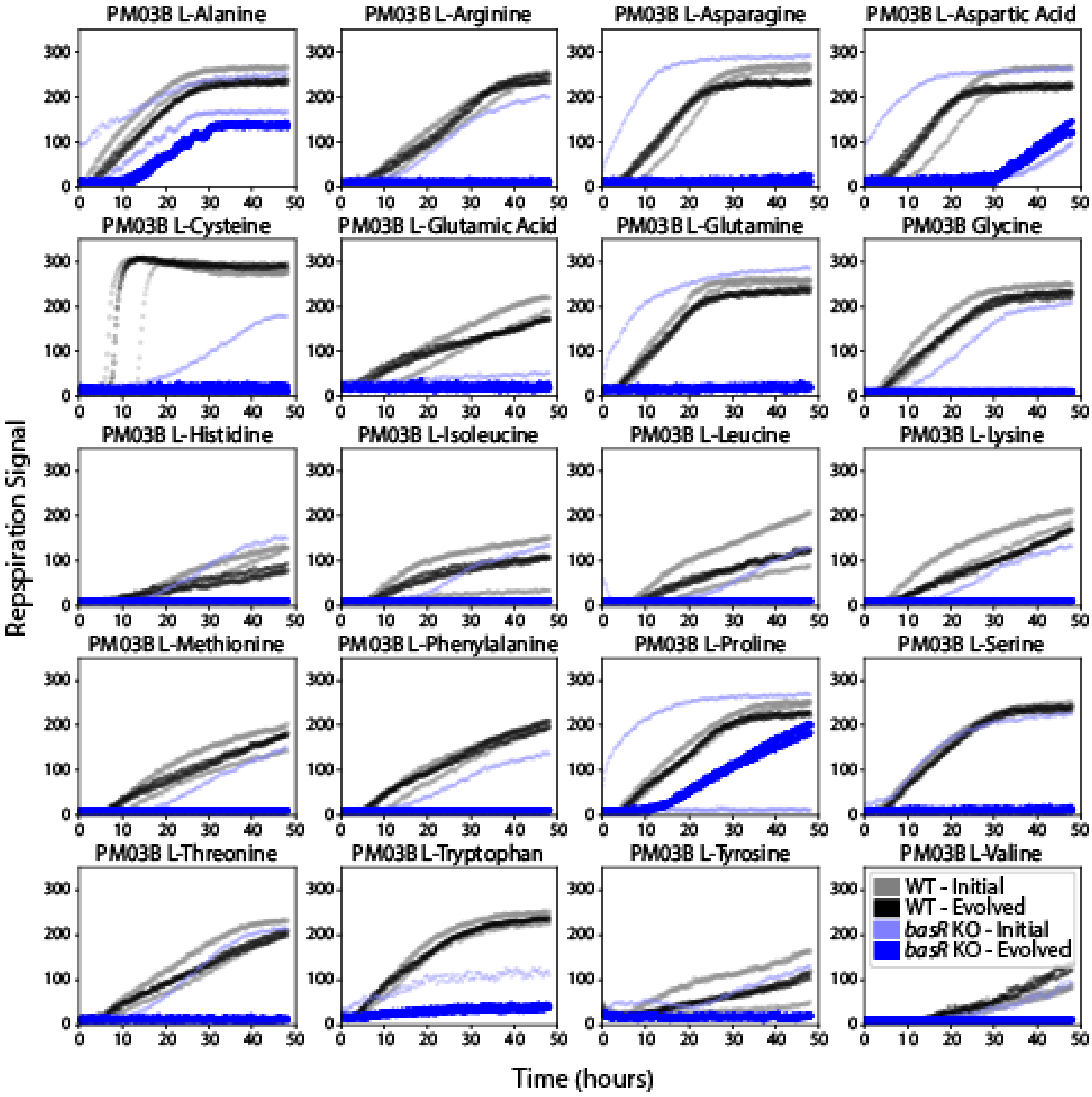

Figure B.7: Respiration signals is an opacity reading that is the measurement for growth for OmniLog plates. Each plot here represents a different nitrogen-limited plate supplemented with an amino acid.

Appendix C

Data-driven modeling of bacterial transcriptional regulation - Supplementary Information

C.1 Supplemental Figures

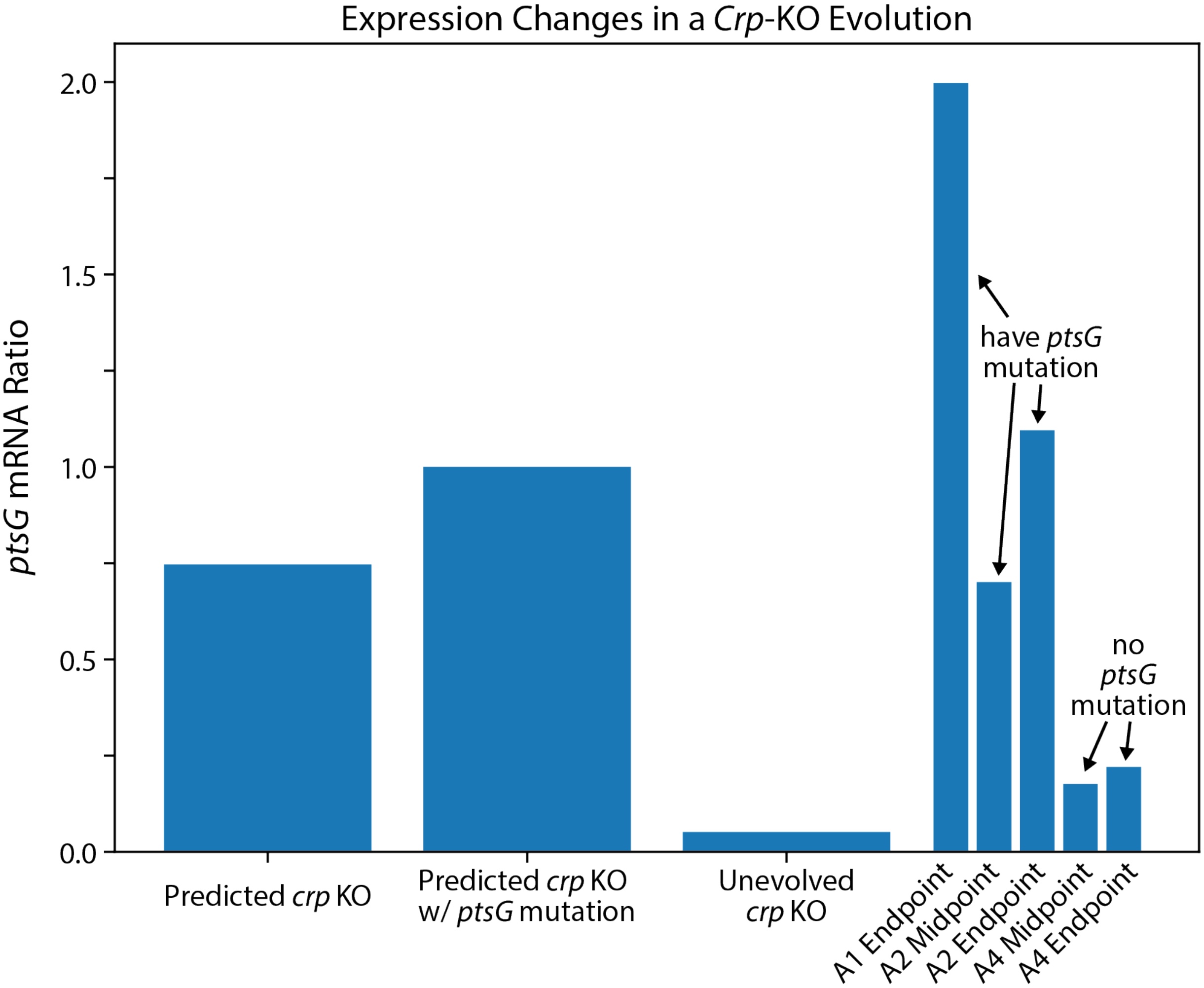

Figure C.1: A *crp* KO evolution study led to convergent mutations of repressor sites upstream of *ptsG*. Our model predicts higher expression with an upstream *ptsG* mutation. A similar trend can be seen in the experiment's data.

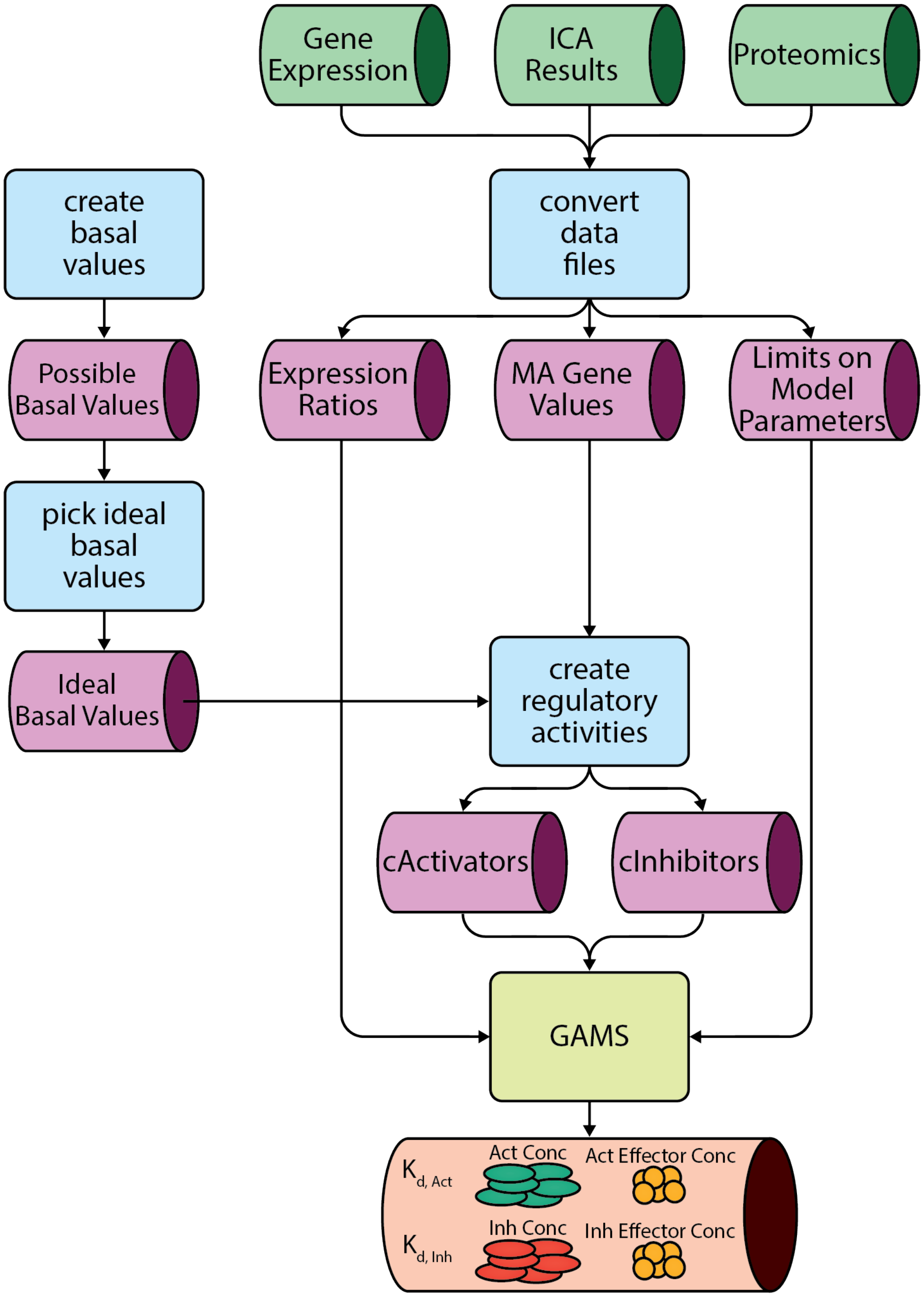

Figure C.2: The cylinders represent data while the squares represent code. Green represents input data, purple are internal data types used within the workflow, and red represents output data. Blue code boxes are run for every gene while GAMS is run for all genes of a regulatory case.

Bibliography

[1] Lee DJ, Minchin SD, Busby SJW (2012) Activating transcription in bacteria. Annu Rev Microbiol 66: 125–152.

[2] Caldara M, Charlier D, Cunin R (2006) The arginine regulon of escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. Microbiology 152: 3343–3354.

[3] Jishage M, Ishihama A (1998) A stationary phase protein in escherichia coli with binding activity to the major sigma subunit of RNA polymerase. Proc Natl Acad Sci U S A 95: 4953–4958.

[4] Hidalgo D, Martínez-Ortiz CA, Palsson BO, Jiménez JI, Utrilla J (2022) Regulatory perturbations of ribosome allocation in bacteria reshape the growth proteome with a trade-off in adaptation capacity. iScience 25: 103879.

[5] Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO (2021) iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. Nucleic Acids Res 49: D112–D120.

[6] Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, Maloof JN, Sinha NR (2012) A High-Throughput method for illumina RNA-Seq library preparation. Front Plant Sci 3: 28988.

[7] UC rate — UC davis and other UC campuses. `https://dnatech.genomecenter.ucdavis.edu/uc-prices/`. Accessed: 2024-5-2.

[8] Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C (2022) The sequence read archive: a decade more of explosive growth. Nucleic Acids Res 50: D387–D390.

[9] Karp PD, Paley S, Caspi R, Kothari A, Krummenacker M, Midford PE, Moore LR, Subhraveti P, Gama-Castro S, Tierrafria VH, Lara P, Muñiz-Rascado L, Bonavides-Martinez C, Santos-Zavaleta A, Mackie A, Sun G, Ahn-Horst TA, Choi H, Covert MW, Collado-Vides J, Paulsen I (2023) The EcoCyc database (2023). EcoSal Plus .

[10] Huerta AM, Salgado H, Thieffry D, Collado-Vides J (1998) RegulonDB: A database on transcriptional regulation in escherichia coli. Nucleic Acids Res 26: 55–59.

[11] Salgado H, Gama-Castro S, Lara P, Mejia-Almonte C, Alarcón-Carranza G, López-Almazo AG, Betancourt-Figueroa F, Peña-Loredo P, Alquicira-Hernández S, Ledezma-Tejeida D, Arizmendi-Zagal L, Mendez-Hernandez F, Diaz-Gomez AK, Ochoa-Praxedis E, Muñiz-Rascado LJ, García-Sotelo JS, Flores-Gallegos FA, Gómez L, Bonavides-Martínez C, del Moral-Chávez VM, Hernández-Alvarez AJ, Santos-Zavaleta A, Capella-Gutierrez S, Gelpi JL, Collado-Vides J (2023) RegulonDB v12.0: a comprehensive resource of transcriptional regulation in e. coli K-12. Nucleic Acids Res 52: D255–D264.

[12] UniProt Consortium (2023) UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 51: D523–D531.

[13] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M (2022) KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 51: D587–D592.

[14] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

[15] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 51: D638–D646.

[16] Phaneuf PV, Gosting D, Palsson BO, Feist AM (2019) ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Res 47: D1164–D1171.

[17] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3: 160018.

[18] Gilbert W, Müller-Hill B (1967) The lac operator is DNA. Proc Natl Acad Sci U S A 58: 2415–2421.

[19] Malan TP, Kolb A, Buc H, McClure WR (1984) Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. J Mol Biol 180: 881–909.

[20] Harden TT, Herlambang KS, Chamberlain M, Lalanne JB, Wells CD, Li GW, Landick R, Hochschild A, Kondev J, Gelles J (2020) Alternative transcription cycle for bacterial RNA polymerase. Nat Commun 11: 448.

[21] Kazmierczak MJ, Wiedmann M, Boor KJ (2005) Alternative sigma factors and their roles in bacterial virulence. Microbiol Mol Biol Rev 69: 527–543.

[22] Roberts JW, Shankar S, Filter JJ (2008) RNA polymerase elongation factors. Annu Rev Microbiol 62: 211–233.

[23] Landick R, Carey J, Yanofsky C (1985) Translation activates the paused transcription complex and restores transcription of the trp operon leader region. Proc Natl Acad Sci U S A 82: 4663–4667.

[24] Conrad TM, Frazier M, Joyce AR, Cho BK, Knight EM, Lewis NE, Landick R, Palsson BØ (2010) RNA polymerase mutants found through adaptive evolution reprogram escherichia coli for optimal growth in minimal media. Proc Natl Acad Sci U S A 107: 20500–20505.

[25] (2022) Prokaryotic ncRNAs: Master regulators of gene expression. Current Research in Pharmacology and Drug Discovery 3: 100136.

[26] Van Assche E, Van Puyvelde S, Vanderleyden J, Steenackers HP (2015) RNA-binding proteins involved in post-transcriptional regulation in bacteria. Front Microbiol 6: 141.

[27] Tollerson R 2nd, Ibba M (2020) Translational regulation of environmental adaptation in bacteria. J Biol Chem 295: 10434–10445.

[28] Gottesman S (2019) Trouble is coming: Signaling pathways that regulate general stress responses in bacteria. J Biol Chem 294: 11685–11700.

[29] Goelzer A, Fromion V (2011) Bacterial growth rate reflects a bottleneck in resource allocation. Biochim Biophys Acta 1810: 978–988.

[30] Utrilla J, O'Brien EJ, Chen K, McCloskey D, Cheung J, Wang H, Armenta-Medina D, Feist AM, Palsson BO (2016) Global rebalancing of cellular resources by pleiotropic point mutations illustrates a multi-scale mechanism of adaptive evolution. Cell Syst 2: 260–271.

[31] Ganesan V, Spagnuolo M, Agrawal A, Smith S, Gao D, Blenner M (2019) Advances and opportunities in gene editing and gene regulation technology for yarrowia lipolytica. Microb Cell Fact 18: 208.

[32] Yu J, Li M, Wang J, Hamushan M, Jiang F, Wang B, Hu Y, Han P, Tang J, Guo G, Shen H (2023) Identification of virulence-modulating RNA from transcriptomics data with machine learning. Virulence 14: 2228657.

[33] Tierrafría VH, Rioualen C, Salgado H, Lara P, Gama-Castro S, Lally P, Gómez-Romero L, Peña-Loredo P, López-Almazo AG, Alarcón-Carranza G, Betancourt-Figueroa F, Alquicira-Hernández S, Enrique Polanco-Morelos J, García-Sotelo J, Gaytan-Nuñez E, Méndez-Cruz CF, Muñiz LJ, Bonavides-Martínez C, Moreno-Hagelsieb G, Galagan JE, Wade JT, Collado-Vides J (2022). RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in escherichia coli K-12.

[34] Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO. A multi-scale transcriptional regulatory network knowledge base for*Escherichia coli.*

[35] Robinson MD, McCarthy DJ, Smyth GK (2009) edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.

[36] Saelens W, Cannoodt R, Saeys Y (2018) A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9: 1090.

[37] Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, Choudhary KS, Yang L, King ZA, Palsson BO (2019) The escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun 10: 5536.

[38] Scott M, Hwa T (2011) Bacterial growth laws and their applications. Curr Opin Biotechnol 22: 559–565.

[39] Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. Science 330: 1099–1102.

[40] Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R (2015) Proteome reallocation in escherichia coli with increasing specific growth rate. Mol Biosyst 11: 1184–1193.

[41] Tan J, Sastry AV, Fremming KS, Bjørn SP, Hoffmeyer A, Seo S, Voldborg BG, Palsson BO (2020) Independent component analysis of e. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. Metab Eng 61: 360–368.

[42] Anand A, Chen K, Catoiu E, Sastry AV, Olson CA, Sandberg TE, Seif Y, Xu S, Szubin R, Yang L, Feist AM, Palsson BO (2020) OxyR is a convergent target for mutations acquired during adaptation to oxidative Stress-Prone metabolic states. Mol Biol Evol 37: 660–667.

[43] Sastry A, Dillon N, Poudel S, Hefner Y, Xu S, Szubin R, Feist A, Nizet V, Palsson B. Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility.

[44] Kim J, Darlington A, Salvador M, Utrilla J, Jiménez JI (2020) Trade-offs between gene expression, growth and phenotypic diversity in microbial populations. Curr Opin Biotechnol 62: 29–37.

[45] Irving SE, Choudhury NR, Corrigan RM (2021) The stringent response and physiological roles of (pp)pgpp in bacteria. Nat Rev Microbiol 19: 256–271.

[46] Schellhorn HE (2020). Function, evolution, and composition of the RpoS regulon in escherichia coli.

[47] Nyström T (2004) MicroReview: Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition? Mol Microbiol 54: 855–862.

[48] Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD (2006) Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell 127: 941–954.

[49] Jovanovic M, Burrows PC, Bose D, Cámara B, Wiesler S, Zhang X, Wigneshweraraj S, Weinzierl ROJ, Buck M (2011) Activity map of the escherichia coli RNA polymerase bridge helix. J Biol Chem 286: 14469–14479.

[50] Ederth J, Artsimovitch I, Isaksson LA, Landick R (2002) The downstream DNA jaw of bacterial RNA polymerase facilitates both transcriptional initiation and pausing. J Biol Chem 277: 37456–37463.

[51] Ross W, Sanchez-Vazquez P, Chen AY, Lee JH, Burgos HL, Gourse RL (2016). ppgpp binding to a site at the RNAP-DksA interface accounts for its dramatic effects on transcription initiation during the stringent response.

[52] Deighan P, Diez CM, Leibman M, Hochschild A, Nickels BE (2008) The bacteriophage lambda Q antiterminator protein contacts the beta-flap domain of RNA polymerase. Proc Natl Acad Sci U S A 105: 15305–15310.

[53] Jin DJ, Cashel M, Friedman DI, Nakamura Y, Walter WA, Gross CA (1988) Effects of rifampicin resistant rpob mutations on antitermination and interaction with nusa in escherichia coli. J Mol Biol 204: 247–261.

[54] Parshin A, Shiver AL, Lee J, Ozerova M, Schneidman-Duhovny D, Gross CA, Borukhov S (2015). DksA regulates RNA polymerase in *Escherichia coli* through a network of interactions in the secondary channel that includes sequence insertion 1.

[55] Zhou Y, Zhang X, Ebright RH (1993). Identification of the activating region of catabolite gene activator protein (CAP): isolation and characterization of mutants of CAP specifically defective in transcription activation.

[56] Rhodius VA, Busby SJ (2000) Transcription activation by the escherichia coli cyclic AMP receptor protein: determinants within activating region 3. J Mol Biol 299: 295–310.

[57] Wytock TP, Fiebig A, Willett JW, Herrou J, Fergin A, Motter AE, Crosson S (2018) Experimental evolution of diverse escherichia coli metabolic mutants identifies genetic loci for convergent adaptation of growth rate. PLoS Genet 14: e1007284.

[58] Sandberg TE, Szubin R, Phaneuf PV, Palsson BO (2020) Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. Nat Ecol Evol 4: 1402–1409.

[59] Radi MS, SalcedoSora JE, Kim SH, Sudarsan S, Sastry AV, Kell DB, Herrgård MJ, Feist AM (2022) Membrane transporter identification and modulation via adaptive laboratory evolution. Metab Eng 72: 376–390.

[60] Chen Y, Boggess EE, Ocasio ER, Warner A, Kerns L, Drapal V, Gossling C, Ross W, Gourse RL, Shao Z, Dickerson J, Mansell TJ, Jarboe LR (2020) Reverse engineering of fatty acid-tolerant escherichia coli identifies design strategies for robust microbial cell factories. Metab Eng 61: 120–130.

[61] Chen J, Gopalkrishnan S, Chiu C, Chen AY, Campbell EA, Gourse RL, Ross W, Darst SA (2019) TraR allosterically regulates transcription initiation by altering RNA polymerase conformation. Elife 8.

[62] Chaudhury S, Lyskov S, Gray JJ (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics 26: 689–691.

[63] Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, Palsson BO, Sommer M, Feist AM (2014). Evolution of escherichia coli to 42 °c and subsequent genetic engineering reveals adaptive mechanisms and novel mutations.

[64] Spira B, Ospino K (2020) Diversity in e. coli (p)ppgpp levels and its consequences. Front Microbiol 11: 564096.

[65] Lamoureux CR, Decker KT, Sastry AV, McConn JL, Gao Y, Palsson BO. PRECISE 2.0 - an expanded high-quality RNA-seq compendium for *Escherichia coli* K-12 reveals high-resolution transcriptional regulatory structure.

[66] Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO (2017) Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc Natl Acad Sci U S A 114: 11548–11553.

[67] Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO. A multi-scale transcriptional regulatory network knowledge base for*Escherichia coli.*

[68] Seo SW, Kim D, O'Brien EJ, Szubin R, Palsson BO (2015) Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in escherichia coli. Nat Commun 6: 7970.

[69] Loiseau L, Vergnes A, Ezraty B (2022). Methionine oxidation under anaerobic conditions in *Escherichia coli.*

[70] Deutscher J, Francke C, Postma PW (2006) How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. Microbiol Mol Biol Rev 70: 939–1031.

[71] Latif H, Federowicz S, Ebrahim A, Tarasova J, Szubin R, Utrilla J, Zengler K, Palsson BO (2018) ChIP-exo interrogation of crp, DNA, and RNAP holoenzyme interactions. PLoS One 13: e0197272.

[72] 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. Lab evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance.

[73] Catoiu E, Phaneuf P, Monk J, Palsson BO. Laboratory-acquired mutations fall outside the wild-type alleleome of *Escherichia coli.*

[74] Kavvas ES, Long CP, Sastry A, Poudel S, Antoniewicz MR, Ding Y, Mohamed ET, Szubin R, Monk JM, Feist AM, Palsson BO (2022) Experimental evolution reveals unifying Systems-Level adaptations but diversity in driving genotypes. mSystems : e0016522.

[75] Kavvas ES, Antoniewicz M, Long C, Ding Y, Monk JM, Palsson BO, Feist AM. Laboratory evolution of multiple *E. coli* strains reveals unifying principles of adaptation but diversity in driving genotypes.

[76] Jayaraman R (2008) Bacterial persistence: some new insights into an old phenomenon. J Biosci 33: 795–805.

[77] King T, Ishihama A, Kori A, Ferenci T (2004) A regulatory trade-off as a source of strain variation in the species escherichia coli. J Bacteriol 186: 5614–5620.

[78] Wessely F, Bartl M, Guthke R, Li P, Schuster S, Kaleta C (2011) Optimal regulatory strategies for metabolic pathways in escherichia coli depending on protein costs. Mol Syst Biol 7: 515.

[79] Robinson R (1993) Cost-benefit analysis. BMJ 307: 924–926.

[80] Banks DL, Rios Aliaga JM, Insua DR (2015). Adversarial risk analysis.

[81] Slovic P, Peters E, Finucane ML, Macgregor DG (2005) Affect, risk, and decision making. Health Psychol 24: S35–40.

[82] Sanchez-Vazquez P, Dewey CN, Kitten N, Ross W, Gourse RL (2019) Genome-wide effects on transcription from ppgpp binding to its two sites on RNA polymerase. Proc Natl Acad Sci U S A 116: 8310–8319.

[83] (2013) The magic spot: A ppgpp binding site on e. coli RNA polymerase responsible for regulation of transcription initiation. Mol Cell 50: 420–429.

[84] Patel A, McGrosso D, Hefner Y, Campeau A, Sastry AV, Maurya S, Rychel K, Gonzalez DJ, Palsson BO (2023) Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. bioRxiv .

[85] Sandberg TE, Wise K, Dalldorf C, Szubin R, Feist AM, Glass JI, Palsson B (2022) Adaptive evolution of a minimal organism with a synthetic genome. iScience .

[86] Wannier TM, Nyerges A, Kuchwara HM, Czikkely M, Balogh D, Filsinger GT, Borders NC, Gregg CJ, Lajoie MJ, Rios X, Pál C, Church GM (2020) Improved bacterial recombineering by parallelized protein discovery. Proc Natl Acad Sci U S A 117: 13689.

[87] Zhao D, Yuan S, Xiong B, Sun H, Ye L, Li J, Zhang X, Bi C (2016) Development of a fast and easy method for escherichia coli genome editing with CRISPR/Cas9. Microb Cell Fact 15: 205.

[88] Choe D, Szubin R, Poudel S, Sastry A, Song Y, Lee Y, Cho S, Palsson B, Cho BK (2021) RiboRid: A low cost, advanced, and ultra-efficient method to remove ribosomal RNA for bacterial transcriptomics. PLoS Genet 17: e1009821.

[89] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranović V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao YP, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M (2021) RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 49: D437–D451.

[90] Murakami KS (2015) Structural biology of bacterial RNA polymerase. Biomolecules 5: 848–864.

[91] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585: 357–362.

[92] McKinney W (2017) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. "O'Reilly Media, Inc.".

[93] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat , Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 10 Contributors (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17: 261–272.

[94] Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20: 238.

[95] Balleza E, López-Bojorquez LN, Martínez-Antonio A, Resendis-Antonio O, Lozada-Chávez I, Balderas-Martínez YI, Encarnación S, Collado-Vides J (2009) Regulation by transcription factors in bacteria: beyond description. FEMS Microbiol Rev 33: 133–151.

[96] Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in escherichia coli. Nucleic Acids Res 31: 1234–1244.

[97] Gao Y, Lim HG, Verkler H, Szubin R, Quach D, Rodionova I, Chen K, Yurkovich JT, Cho BK, Palsson BO (2021) Unraveling the functions of uncharacterized transcription factors in escherichia coli using ChIP-exo. Nucleic Acids Res 49: 9696–9710.

[98] Rigali S, Schlicht M, Hoskisson P, Nothaft H, Merzbacher M, Joris B, Titgemeyer F (2004) Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships. Nucleic Acids Res 32: 3418–3426.

[99] Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM (2019) The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. Metab Eng 56: 1–16.

[100] Pal A, Iyer MS, Srinivasan S, Narain Seshasayee AS, Venkatesh KV (2022) Global pleiotropic effects in adaptively evolved lacking CRP reveal molecular mechanisms that define the growth physiology. Open Biol 12: 210206.

[101] Anand A, Olson CA, Sastry AV, Patel A, Szubin R, Yang L, Feist AM, Palsson BO (2021) Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications. Cell Rep 35: 108961.

[102] McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO (2018) Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in escherichia coli pgi knockout evolutions. Appl Environ Microbiol 84.

[103] Charusanti P, Conrad TM, Knight EM, Venkataraman K, Fong NL, Xie B, Gao Y, Palsson BØ (2010) Genetic basis of growth adaptation of escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet 6: e1001186.

[104] Sekowska A, Wendel S, Fischer EC, Nørholm MHH, Danchin A (2016) Generation of mutation hotspots in ageing bacterial colonies. Sci Rep 6: 2.

[105] Campos M, Govers SK, Irnov I, Dobihal GS, Cornet F, Jacobs-Wagner C (2018) Genomewide phenotypic analysis of growth, cell morphogenesis, and cell cycle events in. Mol Syst Biol 14: e7573.

[106] Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO (2023) A multi-scale expression and regulation knowledge base for escherichia coli. Nucleic Acids Res 51: 10176–10193.

[107] Tierrafría VH, Rioualen C, Salgado H, Lara P, Gama-Castro S, Lally P, Gómez-Romero L, Peña-Loredo P, López-Almazo AG, Alarcón-Carranza G, Betancourt-Figueroa F, Alquicira-Hernández S, Polanco-Morelos JE, García-Sotelo J, Gaytan-Nuñez E, Méndez-Cruz CF, Muñiz LJ, Bonavides-Martínez C, Moreno-Hagelsieb G, Galagan JE, Wade JT, Collado-Vides J (2022) RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in K-12. Microb Genom 8.

[108] Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO (2022) A multi-scale transcriptional regulatory network knowledge base for escherichia coli. bioRxiv : 2021.04.08.439047.

[109] Dalldorf C, Rychel K, Szubin R, Hefner Y, Patel A, Zielinski D, Palsson B (2023) The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions. Res Sq .

[110] Sastry AV, Poudel S, Rychel K, Yoo R, Lamoureux CR, Chauhan S, Haiman ZB, Al Bulushi T, Seif Y, Palsson BO (2021) Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. bioRxiv : 2021.07.01.450581.

[111] Gosset G, Zhang Z, Nayyar S, Cuevas WA, Saier MH Jr (2004) Transcriptome analysis of crp-dependent catabolite control of gene expression in escherichia coli. J Bacteriol 186: 3516–3524.

[112] Nobelmann B, Lengeler JW (1996) Molecular analysis of the gat genes from escherichia coli and of their roles in galactitol transport and metabolism. J Bacteriol 178: 6790–6795.

[113] Maki K, Morita T, Otaka H, Aiba H (2010) A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsg mRNA. Mol Microbiol 76: 782–792.

[114] Geissmann TA, Touati D (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. EMBO J 23: 396–405.

[115] Peterman N, Lavi-Itzkovitz A, Levine E (2014) Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. Nucleic Acids Res 42: 12177–12188.

[116] Liu J, Duncan K, Walsh CT (1989) Nucleotide sequence of a cluster of escherichia coli enterobactin biosynthesis genes: identification of enta and purification of its product 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase. J Bacteriol 171: 791–798.

[117] Seo SW, Kim D, Latif H, O'Brien EJ, Szubin R, Palsson BO (2014) Deciphering fur transcriptional regulatory network highlights its complex role beyond iron metabolism in escherichia coli. Nat Commun 5: 4910.

[118] Choudhury A, Gachet B, Dixit Z, Faure R, Gill RT, Tenaillon O (2023) Deep mutational scanning reveals the molecular determinants of RNA polymerase-mediated adaptation and tradeoffs. Nat Commun 14: 6319.

[119] Hung SP, Baldi P, Hatfield GW (2002) Global gene expression profiling in escherichia coli k12. the effects of leucine-responsive regulatory protein. J Biol Chem 277: 40309–40323.

[120] Brinkman AB, Ettema TJG, de Vos WM, van der Oost J (2003) The lrp family of transcriptional regulators. Mol Microbiol 48: 287–294.

[121] Tani TH, Khodursky A, Blumenthal RM, Brown PO, Matthews RG (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. Proc Natl Acad Sci U S A 99: 13471–13476.

[122] Ernsting BR, Atkinson MR, Ninfa AJ, Matthews RG (1992) Characterization of the regulon controlled by the leucine-responsive regulatory protein in escherichia coli. J Bacteriol 174: 1109–1118.

[123] Caldara M, Charlier D, Cunin R (2006) The arginine regulon of escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. Microbiology 152: 3343–3354.

[124] Perini LT, Doherty EA, Werner E, Senear DF (1996) Multiple specific CytR binding sites at the escherichia coli deop2 promoter mediate both cooperative and competitive interactions between CytR and cAMP receptor protein. J Biol Chem 271: 33242–33255.

[125] Shimada T, Tanaka K, Ishihama A (2016) Transcription factor DecR (YbaO) controls detoxification of l-cysteine in escherichia coli. Microbiology 162: 1698–1707.

[126] Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, Lloyd CJ, Gao Y, Yang L, Palsson BO (2017) Global transcriptional regulatory network for escherichia coli robustly connects gene expression to transcription factor activities. Proceedings of the National Academy of Sciences 114: 10286–10291.

[127] Choe D, Lee JH, Yoo M, Hwang S, Sung BH, Cho S, Palsson B, Kim SC, Cho BK (2019) Adaptive laboratory evolution of a genome-reduced escherichia coli. Nat Commun 10: 935.

[128] Kroner GM, Wolfe MB, Freddolino PL (2019) Lrp regulates One-Third of the genome via direct, cooperative, and indirect routes. J Bacteriol 201.

[129] Grimbs A, Klosik DF, Bornholdt S, Hütt MT (2019) A system-wide network reconstruction of gene regulation and metabolism in escherichia coli. PLoS Comput Biol 15: e1006962.

[130] Gupta KR, Kasetty S, Chatterji D (2015) Novel functions of (p)ppgpp and cyclic di-GMP in mycobacterial physiology revealed by phenotype microarray analysis of wild-type and isogenic strains of mycobacterium smegmatis. Appl Environ Microbiol 81: 2571–2578.

[131] Mackie A, Paley S, Keseler IM, Shearer A, Paulsen IT, Karp PD (2014) Addition of escherichia coli K-12 growth observation and gene essentiality data to the EcoCyc database. J Bacteriol 196: 982–988.

[132] Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L (2008) Evolvability and hierarchy in rewired bacterial gene networks. Nature 452: 840–845.

[133] Koubkova-Yu TCT, Chao JC, Leu JY (2018) Heterologous hsp90 promotes phenotypic diversity through network evolution. PLoS Biol 16: e2006450.

[134] Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of escherichia coli K-12 in-frame, single-gene knockout mutants: the keio collection. Mol Syst Biol 2: 2006.0008.

[135] Anand A, Olson CA, Yang L, Sastry AV, Catoiu E, Choudhary KS, Phaneuf PV, Sandberg TE, Xu S, Hefner Y, Szubin R, Feist AM, Palsson BO (2019) Pseudogene repair driven by selection pressure applied in experimental evolution. Nat Microbiol 4: 386–389.

[136] Bervoets I, Charlier D (2019) Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. FEMS Microbiol Rev 43: 304–339.

[137] Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, García-Sotelo JS, Alquicira-Hernández K, Muñiz-Rascado LJ, Peña-Loredo P, Ishida-Gutiérrez C, Velázquez-Ramírez DA, Del Moral-Chávez V, Bonavides-Martínez C, Méndez-Cruz CF, Galagan J, Collado-Vides J (2018) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli K-12. Nucleic Acids Res 47: D212–D220.

[138] Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in escherichia coli. Nucleic Acids Res 31: 1234–1244.

[139] Lamoureux CR, Choudhary KS, King ZA, Sandberg TE, Gao Y, Sastry AV, Phaneuf PV, Choe D, Cho BK, Palsson BO (2020) The bitome: digitized genomic features reveal fundamental genome organization. Nucleic Acids Res 48: 10157–10163.

[140] Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.

[141] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9: 796–804.

[142] Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7: 129–143.

[143] Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ (2009) The transcription unit architecture of the escherichia coli genome. Nat Biotechnol 27: 1043–1049.

[144] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–8.

[145] Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92–96.

[146] Karp PD, Ong WK, Paley S, Billington R, Caspi R, Fulcher C, Kothari A, Krummenacker M, Latendresse M, Midford PE, Subhraveti P, Gama-Castro S, Muñiz-Rascado L, Bonavides-Martinez C, Santos-Zavaleta A, Mackie A, Collado-Vides J, Keseler IM, Paulsen I (2018) The EcoCyc database. EcoSal Plus 8.

[147] Link H, Fuhrer T, Gerosa L, Zamboni N, Sauer U (2015) Real-time metabolome profiling of the metabolic switch between starvation and growth. Nat Methods 12: 1091–1097.

[148] Choi KY, Zalkin H (1992) Structural characterization and corepressor binding of the escherichia coli purine repressor. J Bacteriol 174: 6207–6214.

[149] Hirsch M, Elliott T (2002) Role of ppgpp in rpos stationary-phase regulation in escherichia coli. J Bacteriol 184: 5077–5087.

[150] Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the escherichia coli alternative sigma factor, σE. Proceedings of the National Academy of Sciences 107: 2854–2859.

[151] Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Agmon E, DeFelice MM, Maayan I, Lane K, Spangler RK, Gillies TE, Paull ML, Akhter S, Bray SR, Weaver DS, Keseler IM, Karp PD, Morrison JH, Covert MW (2020) Simultaneous cross-evaluation of heterogeneous datasets via mechanistic simulation. Science 369.

[152] Kotte O, Zaugg JB, Heinemann M (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Syst Biol 6: 355.

[153] Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M (2016) The quantitative and condition-dependent escherichia coli proteome. Nat Biotechnol 34: 104–110.

[154] Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK, Meng J, Durham BP, Shen C, Varaljay VA, Smith CB, Yager PL, Hopkinson BM (2012) Sizing up metatranscriptomics. ISME J 7: 237–243.

[155] Kubitschek HE, Friske JA (1986) Determination of bacterial cell volume with the coulter counter. J Bacteriol 168: 1466–1467.

[156] Milo R, Jorgensen P, Moran U, Weber G, Springer M (2010) BioNumbers–the database of key numbers in molecular and cell biology. Nucleic Acids Res 38: D750–3.

[157] Marc-Andre Gardner CG Marc Parizeau. DEAP. https://dl.acm.org/doi/10.1145/2330784.2330799. Accessed: 2024-5-18.

[158] Zitzler E, Laumanns M, Thiele L (2001) SPEA2: Improving the strength pareto evolutionary algorithm. TIK Report 103.

[159] Gunasekara SM, Hicks MN, Park J, Brooks CL, Serate J, Saunders CV, Grover SK, Goto JJ, Lee JW, Youn H (2015) Directed evolution of the escherichia coli cAMP receptor protein at the cAMP pocket. J Biol Chem 290: 26587–26596.

[160] Lara-Gonzalez S, Dantas Machado AC, Rao S, Napoli AA, Birktoft J, Di Felice R, Rohs R, Lawson CL (2020) The RNA polymerase α subunit recognizes the DNA shape of the upstream promoter element. Biochemistry 59: 4523–4532.