UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-scale analysis of sequence and regulatory information in *Escherichia coli*

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

in

Bioengineering

by

Cameron Robert Lamoureux

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Ludmil Alexandrov
Professor Jeff Hasty
Professor Joe Pogliano
Professor Kun Zhang

2023

The dissertation of Cameron Robert Lamoureux is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To Philomena.

To Astrid.

To Mom and Dad.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I have so many people to deeply thank for their support, guidance, friendship, love, and tutelage throughout my education. I truly stand on the shoulders of giants. I would like to acknowledge these individuals here.

First of all, I would like to thank my doctoral adviser, Professor Bernhard Palsson. He has been an ideal mentor for me. He not only enables pursuit of fascinating research questions in a variety of spaces but also focuses on practical skills development. I am so grateful for his patience, optimism, and encouragement - it has always meant a lot to me to hear my work described as "astonishing" or "exciting" from someone of his stature.

I have also received outstanding mentorship from several senior scientists in the Systems Biology Research Group (SBRG). In particular, Dan Zielinski has been instrumental; he has a unique capability to see the big picture of research project strategy and design while also being immensely helpful with the nitty-gritty details. I would not at all be where I am today without Dan's help - thank you Dan. Anand Sastry also served as an excellent mentor during my first few years; his eager, kind-hearted guidance was infectious. Zak King also provided formative early guidance, especially with his deep software expertise.

I am thankful for the awesome scientists alongside whom I've worked throughout my time in the SBRG. Kevin Rychel is a rockstar - he has been there for support since the beginning, from taking classes and knowing nothing to leading analyses and providing deep insights. Katherine Decker, John Luke McConn, Donghui Choe, Hyungyu Lim, Ye Gao, and Amitesh Anand have been fantastic collaborators and research partners. Sizhe Qiu and Rita Wan have been impressive and seriously helpful undergraduate researchers in much of my later work.

I would not be here without my friends and colleagues at Emerald Cloud Lab. Having

the opportunity to work there with so many accomplished PhD scientists was instructive. I thank Brian Frezza for giving me a chance straight out of undergraduate study. Ben Kline, Paul Zurek, Robert Teed, Hayley Buchman, Guillaume Robichaud and others provided unbeatable camaraderie for me.

I must also acknowledge my invaluable wife, Philomena: the first Dr. Lamoureux. Her love, support, wisdom, scientific acumen, and sheer excellence were core contributors to my completion of this degree. Our daughter Astrid - while her manner of help was unorthodox - nonetheless enabled me to succeed through sheer adorableness. Philomena and Astrid - I love you.

Last but not least, I must deeply acknowledge my parents. They have supported my education in so many different ways from the start. Their love of learning, encouragement, and loving touch were a perfect blend to bring me to this accomplishment. I find it impossible to appropriately capture the magnitude of their role in my success, for which I am forever grateful. Thank you Mom and Dad, I love you.

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

- **CR Lamoureux**, KS Choudhary, ZA King, TE Sandberg, Y Gao, AV Sastry, PV Phaneuf, D Choe, BK Cho, and BO Palsson. 2020. "The Bitome: digitized genomic features reveal fundamental genome organization." *Nucleic acids research*, 48(8):10157-10163. The dissertation author was the primary author.

Chapter 3 in part is a reprint of material submitted for publication in *Nucleic Acids Research Genomics and Bioinformatics*:

- **CR Lamoureux**, PV Phaneuf, BO Palsson, and DC Zielinski. 2023. "*Escherichia coli* functional non-coding regions are highly conserved." The dissertation author was the primary author.

Chapter 4 in part is a reprint of material published in:

- **CR Lamoureux**, KT Decker, AV Sastry, K Rychel, Y Gao, JL McConn, DC Zielinski, and BO Palsson. 2023. "A multi-scale expression and regulation knowledge base for *Escherichia coli*." *Nucleic acids research*, gkad750. The dissertation author was the primary author.

VITA

2015	Bachelor of Science in Molecular Biophysics and Biochemistry, Yale University
2020	Master of Science in Bioengineering, University of California San Diego
2023	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

CR Lamoureux, KT Decker, AV Sastry, K Rychel, Y Gao, JL McConn, DC Zielinski, and BO Palsson. 2023. "A multi-scale expression and regulation knowledge base for *Escherichia coli*." *Nucleic acids research*, gkad750.

CR Lamoureux, KS Choudhary, ZA King, TE Sandberg, Y Gao, AV Sastry, PV Phaneuf, D Choe, BK Cho, and BO Palsson. 2020. "The Bitome: digitized genomic features reveal fundamental genome organization." *Nucleic Acids Research*, 48(18):10157-10163.

JL McConn, **CR Lamoureux**, S Poudel, BO Palsson, and AV Sastry. 2021. "Optimal dimensionality selection for independent component analysis of transcriptomic data." *BMC Bioinformatics*, 22(1):1-13.

SM Chauhan, S Poudel, K Rychel, **C Lamoureux**, R Yoo, T Al Bulushi, Y Yuan, BO Palsson, and AV Sastry. 2021. "Machine learning uncovers a data-driven transcriptional regulatory network for the crenarchaeal thermoacidophile Sulfolobus acidocaldarius." *Frontiers in Microbiology*, 12:753521.

A Anand, A Patel, K Chen, CA Olson, PV Phaneuf, **C Lamoureux**, Y Hefner, R Szubin, AM Feist, and BO Palsson. 2022. "Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity." *Nature Communications*, 13(1):3682.

R Yoo, K Rychel, S Poudel, T Al-Bulushi, Y Yuan, S Chauhan, **C Lamoureux**, BO Palsson, and A Sastry. 2022. "Machine learning of all Mycobacterium tuberculosis H37Rv RNA-seq data reveals a structured interplay between metabolism, stress response, and infection." *Msphere*, 7(2):e00033-22.

ABSTRACT OF THE DISSERTATION

Multi-scale analysis of sequence and regulatory information in *Escherichia coli*

by

Cameron Robert Lamoureux

Doctor of Philosophy in Bioengineering

University of California San Diego, 2023

Professor Bernhard Ø. Palsson, Chair

Biological information is encoded and transmitted by nucleic acids. Next-generation sequencing technologies have unleashed a flood of large-scale genomics and transcriptomics data capturing this information flow. Here, we develop three analytical frameworks for deriving biological knowledge from this data at multiple scales, using *Escherichia coli* as a model. First, we introduce the Bitome, a single-base-pair resolution representation of genome annotation information for a genome sequence. This binarized construct highlights the uneven patterning of genomic information. Moreover, we leverage this information representation to classify genes based on

adaptive mutability and to quantitatively predict mRNA transcript levels based on promoter sequence. Next, we analyze sequence variation in non-coding regions across 2,350 *E. coli* strains. We demonstrate that annotated functional non-coding features are significantly conserved. We also highlight the sufficiency of non-coding alleles to segment phylogroups, and contrast adaptive mutations with wild-type variation. Finally, we construct a high-precision, single-protocol 1,035-sample RNA-seq compendium called PRECISE-1K. Using unsupervised machine learning, we extract 201 independently-modulated groups of genes (iModulons) that capture the majority of the known transcriptional regulatory network. iModulons also reveal novel regulons and uncover a binding-site basis for different functional behavior within the same regulon. In combination, this expression and regulatory information constitute a knowledge base that may be applied towards the analysis of new data. As a whole, this work introduces a multi-scale suite of analytical tools that enable study of information flow by converting big data to biological knowledge.

Chapter 1

Biology as a big data discipline

The genome is biology's central information repository. It is the blueprint for an organism's structure and function. A single genome provides a static snapshot of a specific organism's DNA sequence; however, genomes are in fact dynamic and may be studied at multiple scales through both space and time. For example, as the central dogma [1] dictates, genetic information flows from DNA to RNA via transcription (Fig. 1.1). This critical process enables an organism to respond to its environment, altering the balance of genomic information available for conversion into the proteins that carry out the actions of life. Indeed, the central dogma relies on the multiple layers of information that are encoded by a genome sequence, from the codons that represent amino acids to the binding sites that enable regulation of transcription. Moreover, genome sequences vary even within a species, owing to the inexorable action of molecular evolution.

Advances in nucleic acid sequencing in the past three decades have revolutionized the study of genomes and their dynamics across these different dimensions and scales. Microbial genomes have accumulated rapidly since the first published complete genome sequence - for

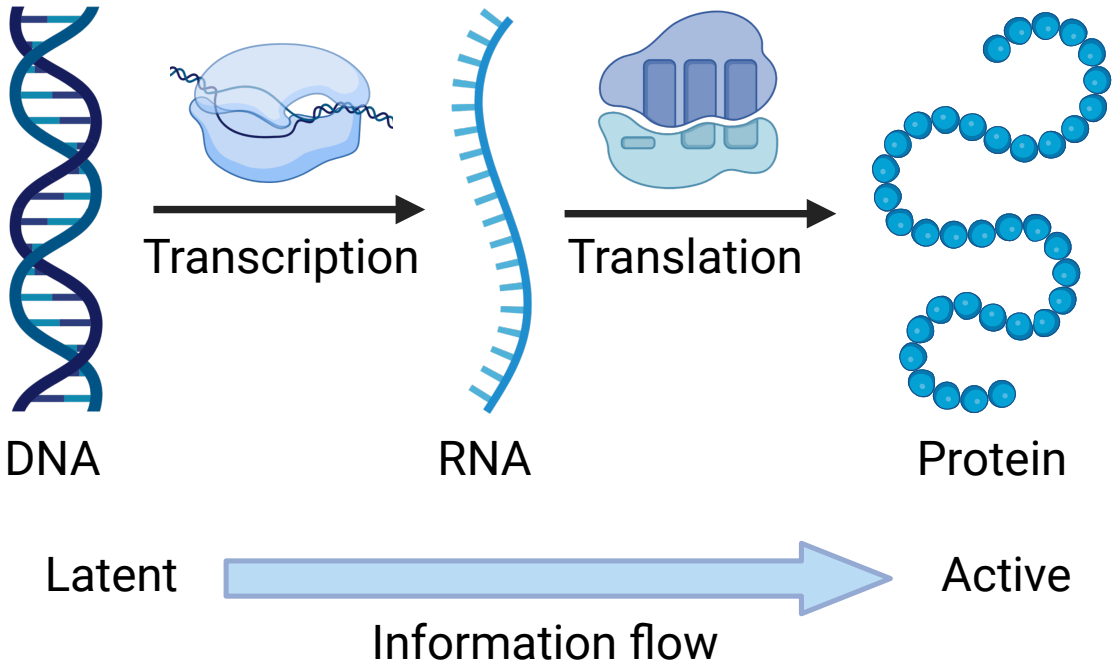


Figure 1.1: The central dogma of molecular biology.

bacterium *Haemophilus influenzae* in 1995 [2]. In particular, sequence data for the model bacterium *Escherichia coli* is plentiful. *E. coli* is of broad importance, owing to its use in the study of: pathogenesis [3–5] and antimicrobial resistance [6, 7]; synthetic biology and the engineering of microbial strains [8–10]; evolution [11]; metabolic modeling [12, 13]; and many more. Over 50,000 *E. coli* genomes are available from a popular online data resource [14] (Fig. 1.2A). Next-generation high-throughput sequencing has enabled the rise of RNA-seq [15], a powerful tool for assessing genome-wide transcriptomic changes in response to environmental stimuli. Indeed, over 10,000 well-annotated *E. coli* transcriptomes are now available on NCBI GEO [16] (Fig. 1.2B). Regulatory interactions and structure in *E. coli* have been elucidated through a combination of decades-long bottom-up biochemical efforts and recent high-throughput advances in chromatin

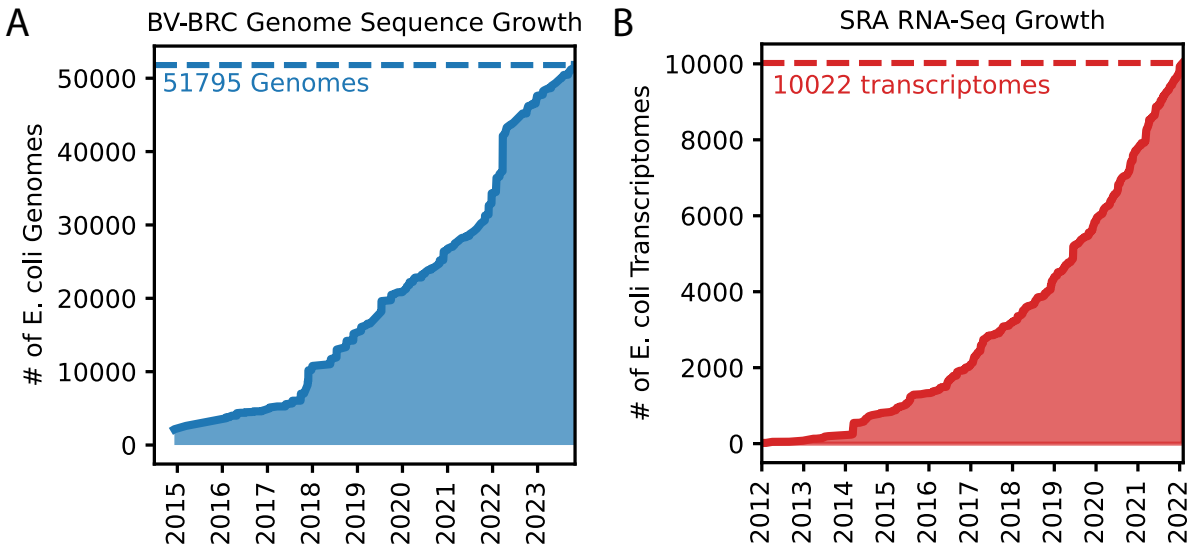


Figure 1.2: Growth in sequencing data. **A)** Growth in genome sequences for *Escherichia coli* deposited in the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). **B)** Growth in transcriptomes for *E. coli* deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).

immunoprecipitation (ChIP) [17, 18], with a reference database [19] containing over 8,000 interactions.

Converting this wealth of data to biological knowledge is a fundamental goal of twenty-first century computational biology [20]. Integrating these different vantage points of *E. coli*'s genome information will enrich understanding of the relationship between genotype and phenotype. Biological data analysis methods that derive new insights from existing data will address the "reusability" aspect of the FAIR data management guiding principles [21].

1.1 Genome annotation at scale

A genome sequence explicitly contains the nucleotide that make up an organism's chromosomes. However, implicit in the sequence are additional layers of information that are represented by different nucleotide groupings. Annotating these different components of a genome is a critical

task to which significant effort has been dedicated. Reference sequence annotations highlight the locations of core features such as coding genes, non-coding RNAs, and mobile elements. Coding region nucleotides are by definition components of one of 64 codons specifying a protein's amino acid sequence. Thus, coding region nucleotides also implicitly contain the information dictating a protein's structure; indeed, recent advances in machine learning have unlocked this relationship to an astonishing degree [22].

Cis-regulatory genomic regions facilitate the activation of genomic information via transcription. These regions comprise a number of distinct types of information that may be mapped to specific nucleotides. Transcription start sites (TSS) and transcription termination sites (TTS) are critical bases that delineate the portions of DNA that are converted into mRNA; TSS, TSS, and the transcriptional units (TUs) they define have been identified at scale in *E. coli* [23]. Core promoter regions control RNA polymerase binding via sigma factor recognition [24–27]. Transcription factors (TFs) recognize specific transcription factor binding sites (TFBS) within *cis*-regulatory regions and activate or repress transcription of the downstream genes in response to environmental cues [19, 28–30]. Transcriptional attenuators represent yet another grouping of nucleotides that contain an additional information layer - in this case on the basis of the mRNA secondary structures they encode [31]. Overall, these layers of genome annotation represent the information encoded by each individual nucleotide in a genome sequence. Unifying this information in a formal, actionable construct remains a desirable next step.

1.2 Genome sequences at scale

The huge scale of genome sequences available for *E. coli* and other organisms has necessitated the development of pangenomics [32]. Pangenomic analyses aim to convert thousands

of individual genome sequences from a species into biological knowledge. Central to this framework are the concepts of core (conserved) and accessory (variable) genomes, which have been defined for many significant microbial species [33]. These demarcations consist of genes found across nearly all genomes in a pangenomic population (core) or found in a non-negligible subset (accessory). Pangenome analysis hinges fundamentally on the clustering of gene nucleotide sequences across genomes based on their sequence similarities, followed by binarization of gene presence and absence across the genome compendium. The study of antimicrobial resistance [34], metabolism [35], and virulence [36] have all been enriched by pangenomic analysis.

Pangenome analysis is almost by definition focused on coding regions. While coding sequences make up the majority of a typical microbial genome [37], non-coding regions play an outsize role in manifesting phenotype from genotype. As discussed previously, *cis*-regulatory regions in particular are rich with critical information layered at single nucleotide resolution. Indeed, the single-nucleotide scale of genomic information highlights another limitation of pangenomics: pangenomic analyses do not directly take into account sequence variation. Extension of pangenomic analyses to non-coding regions at single nucleotide resolution would enable multi-scale analysis of sequence information and variation across multiple genome sequences.

1.3 Transcriptomics and regulation at scale

The availability of large-scale transriptomic data sheds light on the processes by which *E. coli* responds to environmental stimuli. As the cell is exposed to different growth conditions - media, temperature, pH, carbon sources, aerobicity, etc. - the transcription levels of all genes in the genome are modulated. These modulations are largely mediated by transcription factors: DNA-binding proteins that either activate or repress transcription initiation at a particular promoter

region by binding to a specific *cis*-regulatory site, altering the mRNA level of the transcription unit transcribed from that location. Taken together, these transcription factor-transcription unit interactions comprise a transcriptional regulatory network (TRN). This network is the information processing core of the cell, controlling both the spatial and temporal dissemination of genomic information.

The *E. coli* TRN is relatively well-characterized thanks to approaches such as ChIP [17] paired with differential expression analysis. Indeed, over 75% of *E. coli* genes are represented in RegulonDB [19], the premier source for *E. coli* regulatory network annotation that catalogs and curates information on regulons, or groups of co-regulated genes. Nonetheless, even explicit knowledge of the TRN's connectivity is not sufficient to explain or predict gene expression levels [38]. Thus, analytical methods that can further expand knowledge of the TRN and its dynamics are critical - ideally without necessitating continued laborious bottom-up network characterization.

1.4 Thesis outline

In this thesis, we present a set of three analytical frameworks for interrogating genomic and transcriptomic information at scale (Fig. 1.3). In the next chapter, we introduce the Bitome, a formalized representation of the multiple layers of annotation information for a single genome sequence at single nucleotide resolution. This structure enables analysis of the patterning of genome information and prediction of genomic properties based on information content for the model *E. coli* strain K-12 MG1655. The third chapter widens the analytical scale to thousands of genomes; we construct a non-coding alleleome for *E. coli*. By querying sequence variation within non-coding regions across the *E. coli* pangenome, we reveal patterns of conservation and provide

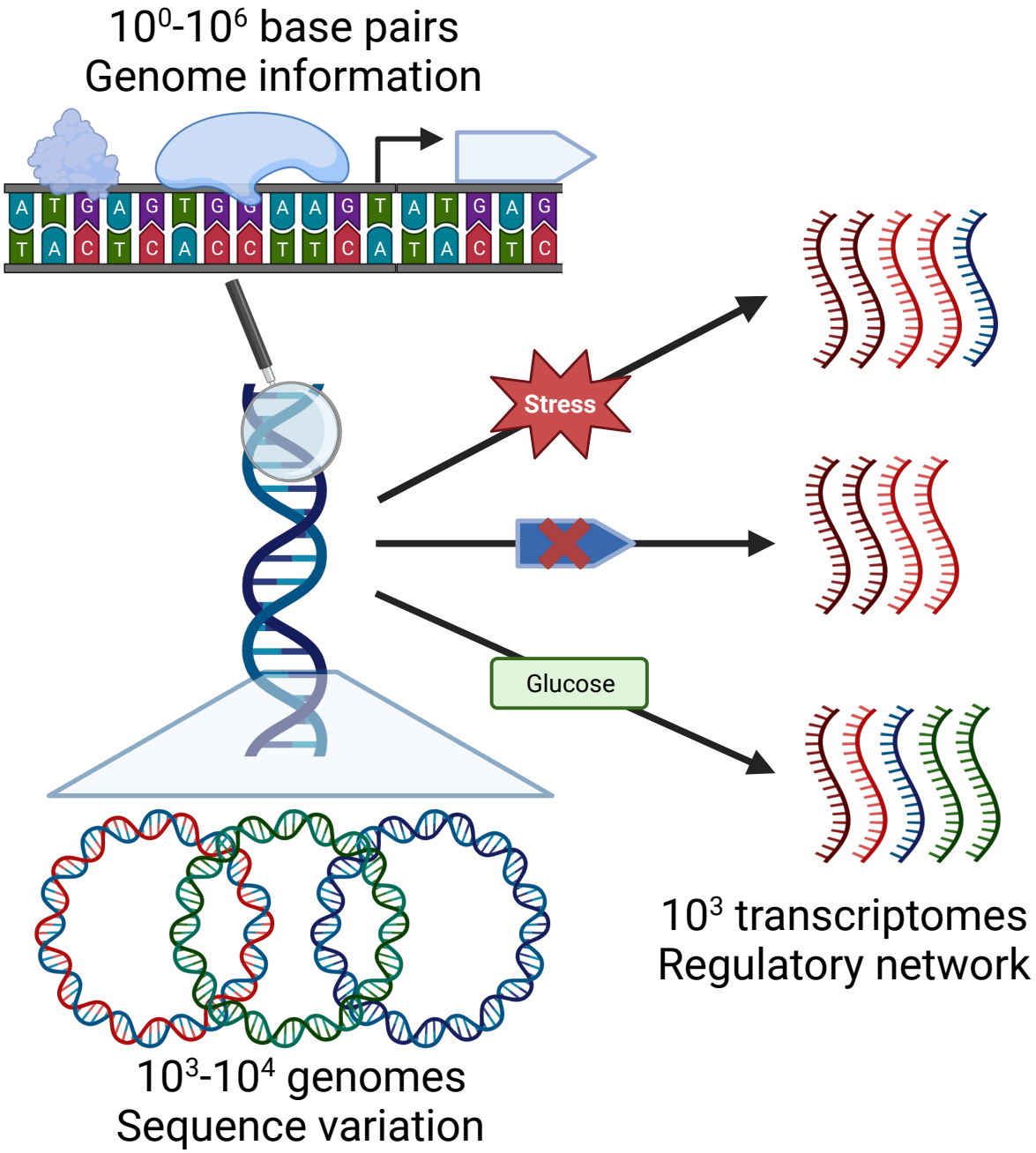


Figure 1.3: Scales of analysis of sequence information.

an additional dimension for the contextualization of genomic information contained in the *E. coli* model strain bitome. Finally, the fourth chapter establishes a large-scale transcriptomics

dataset for *E. coli*, along with a machine learning method for elucidation of regulatory network modules from the dataset. Taken together, these analyses leverage big data to extract actionable biological knowledge regarding the patterning, variation, and processing of genome information at multiple scales.

Chapter 2

The Bitome: Digitized genomic features reveal fundamental genome organization

The information that determines the structure and functioning of an organism is stored in its genome. However, we currently lack a formal framework for representing and studying this information. Here, we introduce the Bitome, which is a matrix consisting of binary digits (bits) that represent the genomic positions of various features within the genome. We construct a Bitome for the genome of *Escherichia coli* K-12 MG1655 and make the following discoveries: (i) genomic features are unevenly encoded, both spatially and categorically; (ii) coding and intergenic features are captured at base-pair resolution; (iii) adaptive mutations tend to occur more frequently in genomic positions with fewer features; and (iv) the Bitome feature information representation empowers classification of both genes with adaptive mutations and essential genes.

The Bitome serves as a formal representation of a genome and offers a valuable tool for studying its fundamental organizational properties.

2.1 Background

A genome contains various types of information that determine an organism's structure and function [39]. Experimental methods at genome scale help uncover genomic features such as sequence [40], transcription units [41], and regulatory elements [42]. This information is vital for endeavors like genome-scale metabolic reconstructions [43,44], characterization of transcriptional regulatory networks [45], and genome design and reduction efforts [8, 46].

However, the current representation of genomic information is predominantly focused on open reading frames and is limited to text or image formats, which hampers comprehensive analysis of genomic information. The genome exhibits structural organization in the form of macrodomains [47], and the location of a gene can influence its expression levels [48]. For instance, the Y-ome, representing *E. coli* genes lacking functional evidence, tends to be concentrated near the terminal region [49]. Fundamental genome properties such as GC content exhibit periodic patterns across different length scales [50]. These findings highlight the need for a formal construct centered around base pairs that can encompass all encoded features of the genome sequence.

To address this need, we introduce the Bitome, a matrix that associates each position in a genome sequence with the corresponding encoded features. As a demonstration, we created a Bitome for the *E. coli* K-12 MG1655 genome. Our observations include: (i) uneven patterning of genomic features throughout the sequence; (ii) differing feature information density within coding and intergenic regions, distinguishing sub-features within them; (iii) a higher frequency of

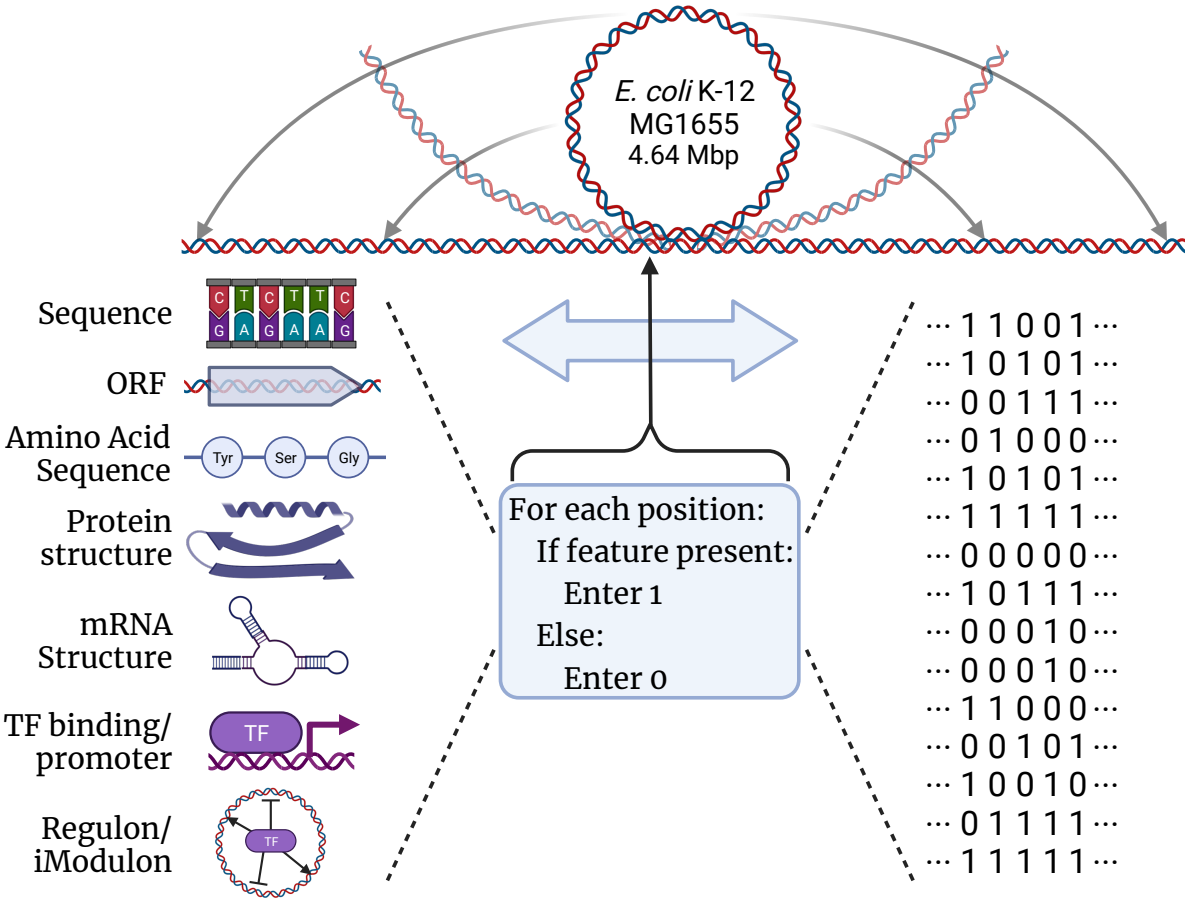


Figure 2.1: Schematic representation of the Bitome. Genomic features are associated with genomic positions in which they appear, enabling binarization of genomic feature information.

adaptive mutations in genomic positions with fewer features; and (iv) the predictive power of the Bitome formalization in identifying adaptively mutated genes and predicting gene essentiality based solely on sequence features. Thus, the Bitome represents a novel construct that formally describes genomic feature information and lays the foundation for actionable predictions based on this information.

2.2 Results

2.2.1 The Bitome formalizes genomic features at base-pair resolution

We generated a Bitome for *E. coli* K-12 MG1655, where each row corresponds to a specific genomic feature, and each column represents a genomic position. The matrix elements, referred to as 'bits', have a binary value of either 1 or 0. A value of 1 indicates the presence of a particular feature i at a given genomic position j, while 0 indicates its absence (Fig. 2.2A). The Bitome encompasses various types of genomic features, including: (a) sequence-derived features, such as codons, (b) experimentally determined features, such as transcription factor binding sites, and (c) computationally predicted features, like protein secondary structure (Fig. 2.2B). The K-12 Bitome consists of 1634 rows representing genomic features and 4,641,652 columns representing genomic positions, encompassing a total of 52.4 million bits. The Bitome is sparse, with only 0.7% of the bits having a value of 1.

2.2.2 Genomic features are patterned unevenly

The structure of the Bitome is exemplified by the *gadAXW* operon. The number of bits in each row of this region varies widely, from the full 4243 bits (indicating the presence of an operon) to 0 bits (for example, most transcription factors don't have binding sites) (Fig. 2.2D). Coding regions exhibit higher bit counts per column compared to intergenic regions (Fig. 2.2E). This difference becomes evident when focusing on a Bitome region located at the edge of a coding gene (Fig. 2.2C). The intergenic regions within this operon are relatively rich in features, containing multiple transcription factor binding sites and tightly structured mRNA secondary structures. The maximum bit count per column is significantly lower than the row dimension of the Bitome, indicating that only a small fraction of the total genomic features have bits at a

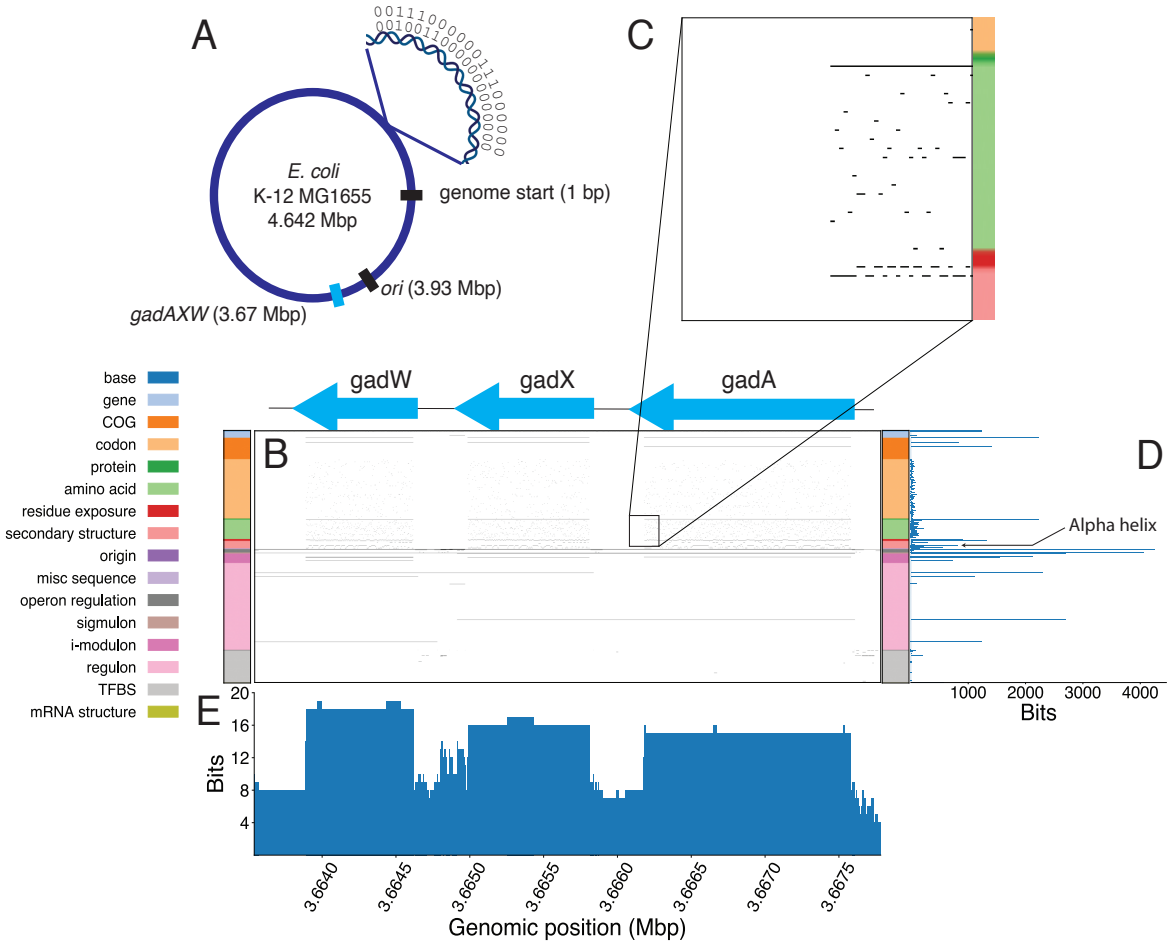


Figure 2.2: Features encoded by the E. coli K-12 MG1655 genome can be represented as a binary matrix. **(A)** *E. coli* K-12 MG1655 genome with reference genome start position, origin of replication (*ori*), and the *gadAXW* operon marked. **(B)** A visualization of the Bitome section at the location of the gadAXW operon. Rows are genomic features, columns genomic position. Black = 1, white = 0. **(C)** Close-up visualization of a 200 x 200 section of the Bitome section in **(B)**. **(D and E)** Bit counts of the rows **(D)** and columns **(E)** of this section.

specific genomic position.

Across the entire Bitome, the bit counts per genomic position range from 2 to 26 (Fig. 2.3A). The majority of positions have between 10 and 15 bits, while a small selection of positions have more than 15 bits. There is notable variation in the percentage of the total sequence that encodes different features. For instance, genomic positions encoding carbon metabolism genes cover 9.1% of the genome, whereas Shine-Dalgarno sequences cover only 0.004%

(Fig. 2.3B). About 35% of the genomic sequence codes for hydrophobic amino acids like leucine, alanine, glycine, valine, and isoleucine (Fig. A.1A). Alpha helices are a common structural motif, encoded by 29% of genomic positions (Fig. A.1B). The organizational structure of the Bitome enables easy computation of sequence coverage for overlapping features. For example, we observed that glycine is more prevalent in loop regions compared to alpha helices or beta sheets (Fig. A.1C). We conducted hierarchical clustering of genomic information within genes, transcription units, and operons, and found that clusters were predominantly influenced by more densely encoded features such as amino acids, without definitively associating features across categories.

2.2.3 Defining coding and intergenic sub-regions by bit density

The density of bits, measured as bits per base pair (bits/bp), varies across different regions of the genome. At a resolution of 100 kb, the moving average of bit density shows fluctuations (Fig. 2.3C). The peak observed at 0.75 Mb is primarily attributed to an increased density of transcription units in that specific region. Notably, the variation in bit density is not distinctly periodic. Furthermore, bit density serves as a distinguishing factor between coding and intergenic features. Protein-coding genes and transcription units generally have a density of 12 bits/bp, while pseudogenes tend to be less feature-rich (Fig. 2.3D).

The Bitome also provides insights into the bit density within intergenic regions. For instance, the 5' and 3' untranslated regions (UTRs) flanking coding genes collectively define a transcription unit (TU) [41]. These regions have a median length of approximately 50 bp but can vary significantly in size (Fig. 2.4A). Intergenic regions within TUs and the 5' and 3' UTRs exhibit an approximate bit density of 6-7 bits/bp (Fig. 2.4B). Overall, when including

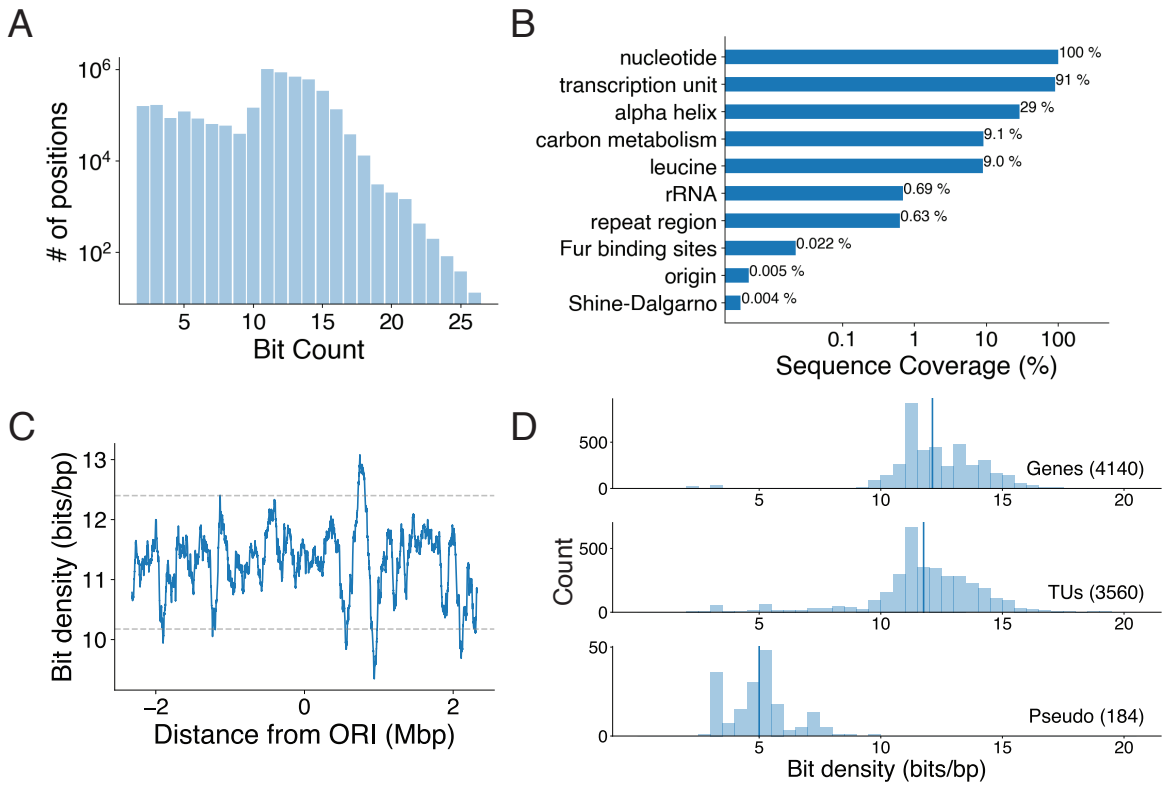


Figure 2.3: Bits are distributed unevenly. **(A)** Histogram of genomic positions by bit count. **(B)** Sequence coverage of 10 selected genomic features. **(C)** Moving average of bit density across the genome, calculated in 100 kb windows. Gray dashed lines indicate the mean ± 2 standard deviations. **(D)** Histograms of bit density for selected features (number of features indicated in parentheses). Vertical lines indicate medians.

these UTRs and intergenic regions within TUs, TUs occupy approximately 91% of the genome sequence (Fig. 2.3B). Consequently, the remaining 9% of the genome consists of "inter-TU" regions.

These inter-TU regions exhibit a deficiency in features, characterized by a median bit density of only 2.5 bits per bp (Fig. 2.4B). These areas encompass transcriptional regulatory sequences such as -10 and -35 elements. Interestingly, the actual positions of these sequences slightly deviate from their designated nomenclature, with the -35 elements, in particular, tending to be located approximately 2 bp closer to the transcription start site (TSS) (Fig. 2.4C). The

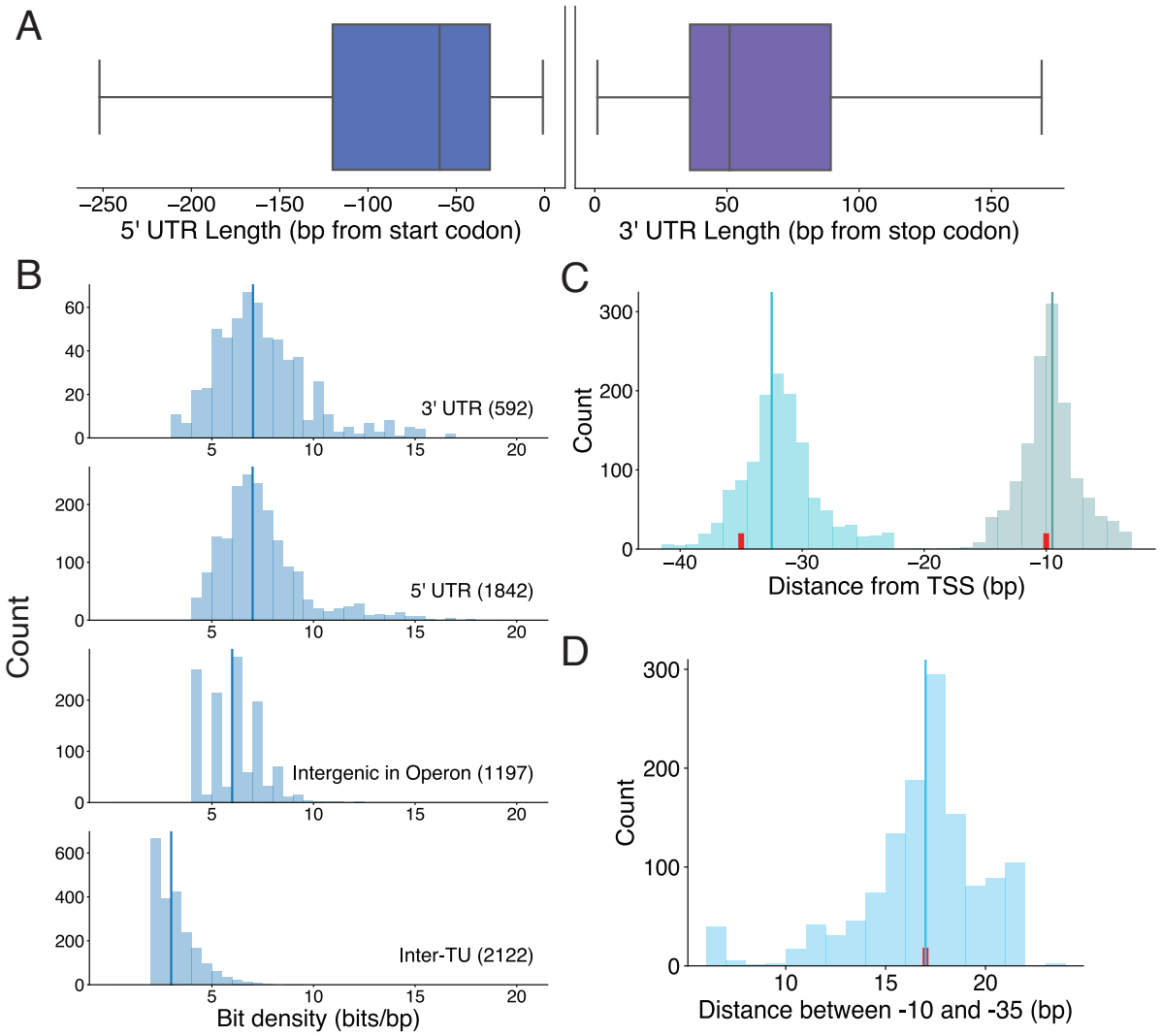


Figure 2.4: The Bitome provides a high-resolution view of bit density in intergenic regions. **(A)** Boxplots of the 5' (blue) and 3' (purple) UTR lengths. 5' UTR: n = 1842, 152 outliers excluded. 3' UTR: n = 594, 94 outliers excluded. Outliers excluded based on 1.5*IQR from Q1 and Q3 (included range indicated by whiskers). **(B)** Histograms of bit density of selected intergenic regions (number of regions indicated in subplot titles). Vertical lines indicate medians. Bits from both strands are considered. **(C)** Histograms of the distributions of the -10 (light green) and -35 (cyan) elements of promoter regions. The center of the element is used to compute distance to TSS. Red ticks indicate the canonical locations of the elements, and vertical lines indicate medians. n = 1306. **(D)** Histogram of distances between -10 and -35 elements from the same promoter (as measured from ends of elements). Red tick indicates literature value. Vertical line indicates median. n = 1306.

distance between these boxes, known to be crucial for RNA polymerase binding to the promoter region [51], is a feature of the inter-TU region accurately represented by the Bitome (Fig. 2.4D). Despite the presence of regulatory sequences, our knowledge of the inter-TU regions is primarily limited to the nucleotide sequences themselves.

2.2.4 Adaptive mutations are biased towards low-information genomic positions

The Bitome provides valuable insights into distal causation during adaptive laboratory evolution (ALE). During ALE experiments, single nucleotide polymorphisms (SNPs) are acquired throughout the genome [52]. Interestingly, coding SNPs occur less frequently at genomic positions that have a higher bit density, particularly in coding regions (Fig. 2.5A). Among the amino acids, threonine, which ranks as the sixth most abundant in terms of sequence coverage, is observed to be the most frequently mutated amino acid, with a mutation frequency surpassing that of the overall sequence (Fig. 2.5B). In contrast, despite leucine being the most abundant amino acid in terms of sequence coverage, it is mutated at only two-thirds of the overall genome mutation rate. Additionally, hydrophobic residues, despite providing a larger sequence target, are less frequently subject to missense mutations.

2.2.5 The Bitome enables prediction of adaptively mutations and gene essentiality

The Bitome is capable of predicting genes that undergo SNP acquisition during ALE. By utilizing only the bits from coding gene regions, we trained a support vector machine (SVM) classifier (Fig. 2.5C) to discern between coding genes that acquire SNPs and those that do not

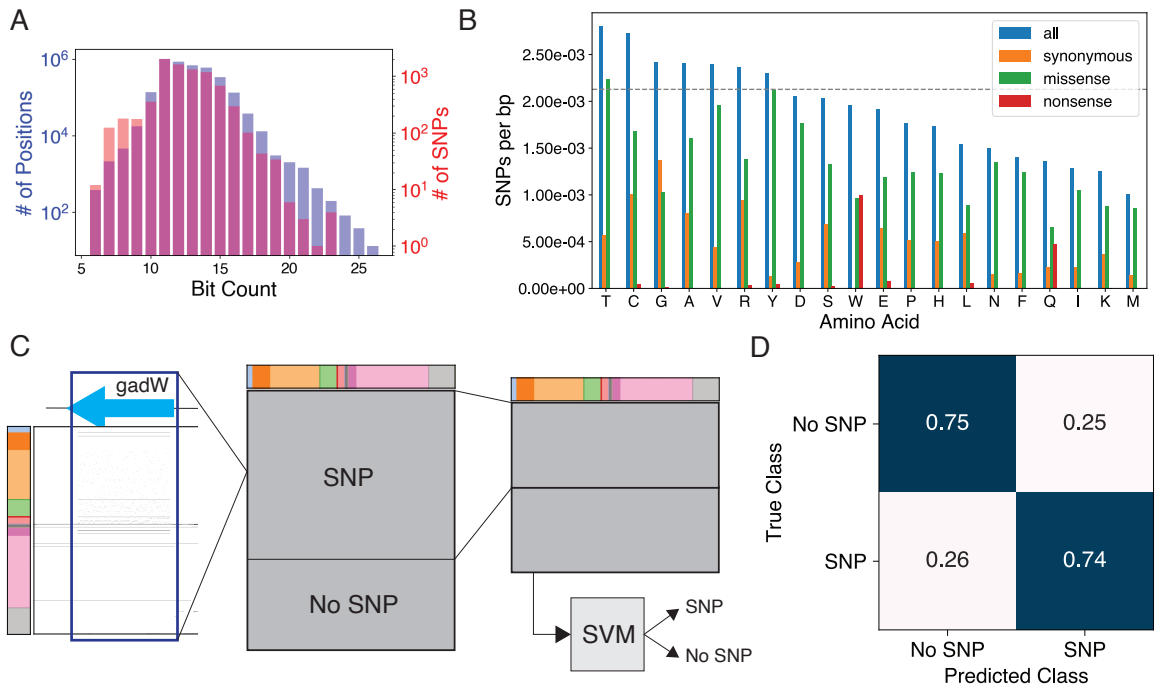


Figure 2.5: The Bitome enriches systemic analysis and prediction of adaptive mutations. **(A)** Combined histogram of the number of coding genome positions that contain the given number of bits (purple) and the numbers of SNPs that occur in coding positions with that number of bits (red). Two-sided Mann-Whitney U test: P = 0.015; n = 3 881 981, m = 7034. **(B)** Frequency of SNPs occurring at each amino acid. The gray dashed line is the overall frequency of SNPs across the entire genome. **(C)** Diagram of pipeline for predicting genes acquiring SNPs during ALE. From left to right: Bitome region for gene summed column-wise to give feature vector. Gene feature vectors combined into gene feature matrix and labeled as having at least one ALE SNP or not. Training matrix constructed by random down-sampling of majority class (SNP). Support vector machine (SVM) model trained to classify genes. Colorbar represents Bitome features as in Fig. 2.2B. **(D)** Confusion matrix for final model. Scores are accuracy, normalized to true class. n = 506 in held-out, lockbox test set.

during ALE experiments. The SVM model achieves an accuracy of 75% ± 1% in this classification task (Fig. A.2B), exhibiting no bias towards any specific class (Fig. 2.5D). Interestingly, even after excluding the nucleotide sequence features, the model maintains its accuracy; however, the performance declines when solely relying on the sequence (Fig. A.1A). Thus, the Bitome faithfully represents valuable genomic information that is encoded by the sequence but not readily deducible from it. Notably, the model identifies the specific stop codon UAG as a significant feature for

predicting genes with observed SNPs, while membership in the sigma factor 32 or Fis/Lrp/H-NS regulons proves important for predicting non-mutated genes (Fig. A.1E).

We applied a similar approach to classify essential genes from the Keio collection [53] using Bitome features. The support vector machine classifier achieved an area under the receiver operating characteristic curve (AUC) of 0.75 (Fig. A.3A), displaying a slight imbalance favoring the non-essential class (Fig. A.3B). Nonetheless, the classifier successfully identified meaningful clusters of orthologous groups (COGs) that are relevant for prediction, such as cell cycle and translation (Fig. A.3D). Notably, residue exposure emerged as an important feature in classifying essentiality, highlighting the Bitome's potential to unveil unexpected connections between genomic features and phenotypic outcomes.

2.2.6 Intergenic sequence-based features enable quantitative prediction of *in vivo* transcript levels

We then leveraged the Bitome's intergenic features to predict *in vivo* transcript levels from a comprehensive *E. coli* expression compendium [54]. We performed exploratory data analysis to investigate each of our sequence-based features and their relationships to transcript levels. We found minimal high correlations between our sequence-based genomic features, indicating that they all had the potential to add useful information to our model. Amongst the local promoter features, the sequence of the -10 element and the nucleotide at the TSS demonstrated notable relationships to expression level. The canonical TATAAT -10 element sequence produced the highest median expression level, and the 1-bp-variant -10 element sequences TAAAAT and TAGAAT followed, amongst well-represented -10 element sequences (Fig. 2.6A). However, the expression variance both within the same -10 element sequence (over three orders of magnitude)

and between the medians of different -10 element sequences (over two orders of magnitude) indicated that this feature - while important - does not suffice to quantitatively distinguish expression levels. The TSS nucleotide follows a similar pattern, with the four different groups showing significantly different expression (one-way F test; P=2.88E-5). Interestingly, the most common TSS nucleotide - A - does not produce the highest median expression, and we see a preference for a GC base pair at the TSS (Fig. 2.6B).

Replication region stood out amongst the genome-scale features as a contributor to differences in expression level throughout the genome. Differences in median expression between genes on the leading strand, on the lagging strand, and in the terminus region were significant (one-way F test; P=3.5E-14) (Fig. 2.6C). Genes on the leading strand were more highly expressed than those on the lagging strand, which in turn exceeded the levels of genes found in the terminus region. Interestingly, this genome-scale feature appeared to affect gene expression levels more than the actual distance to the origin of replication.

We then used our set of sequence-based features to train machine learning models for transcript level prediction. We profiled a range of machine learning models of varying complexities using a five-fold cross validation approach. We found that a random forest model outperformed the alternative models at this prediction task, achieving an R-squared score of 0.64 upon cross-validation (Fig. 2.7A). A support vector machine model also performed adequately, yielding an R-squared score of 0.5, with somewhat less overfitting than the random forest model.

Our quantitative expression model utilized a mixture of expected and surprising features to drive its predictive success. -10 box score, consistent with its notable role at the exploratory stage and well-documented importance for transcription, was the second most important feature in the model, with -35 box score at fifth-most important (Fig. 2.7B). The distance from a gene

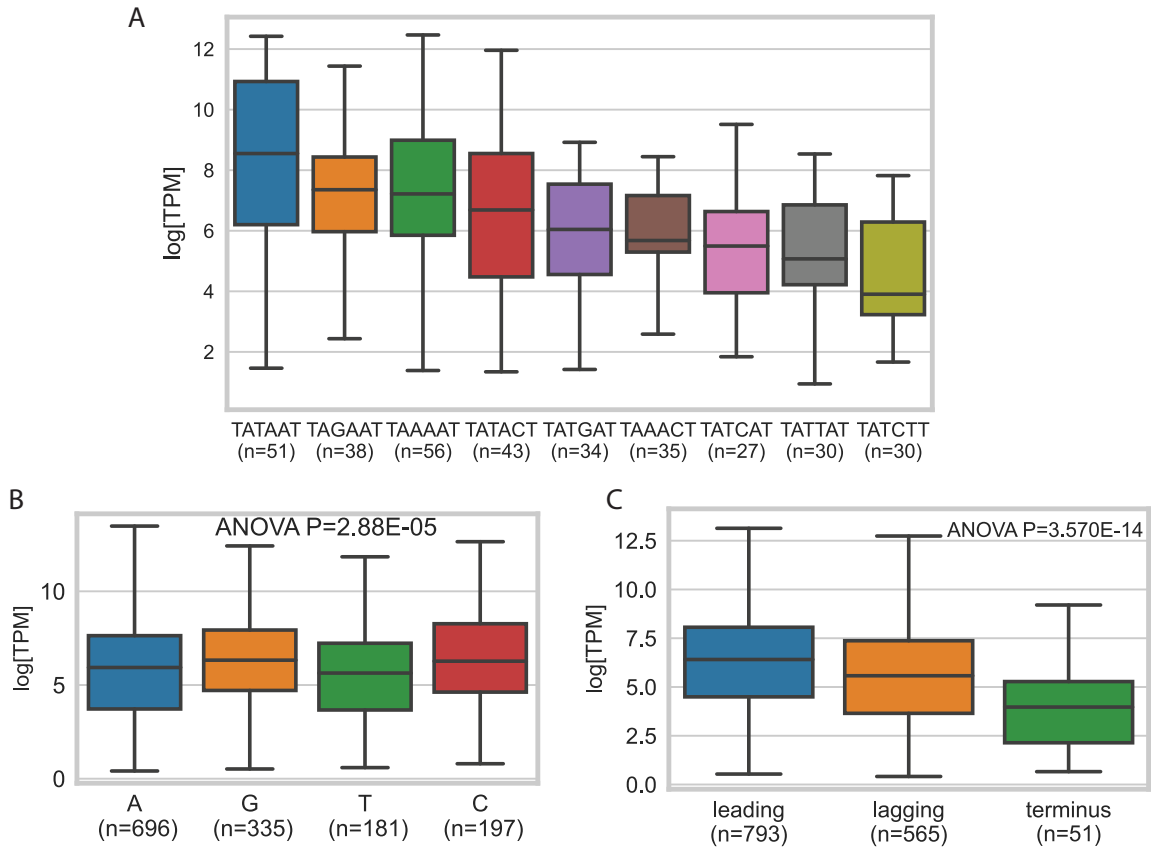


Figure 2.6: Local and genome-scale features distinguish expression levels. **(A)** Distributions of expression levels for genes with particular -10 box sequences (9 box sequences with more than 25 examples shown out of 227 total unique -10 box sequences). Box whiskers indicate 1.5*IQR from Q1 and Q3 (included range indicated by whiskers); center lines are medians. **(B)** Distributions of expression levels for each TSS base. **(A)** Distributions of expression levels for each replication region. Leading: genes transcribed in the same direction as genome replication; lagging: genes transcribed opposite to genome replication; terminus: genes located between the terA and terC sequences for which leading/lagging regions are ill-defined based on stochastic termination of replication by replication fork collision within terminus region.

to the boundaries of experimentally-determined CIDs - also previously implicated as a partial determinant of expression level [56] - did also prove useful to the final model; the model picked out the expected relationship, with lower distances to the CID boundaries yielding higher expression predictions. However, the GC content and length of the 5' UTR, as first and fourth most important features, respectively, stood out from the model. These features did not stand out

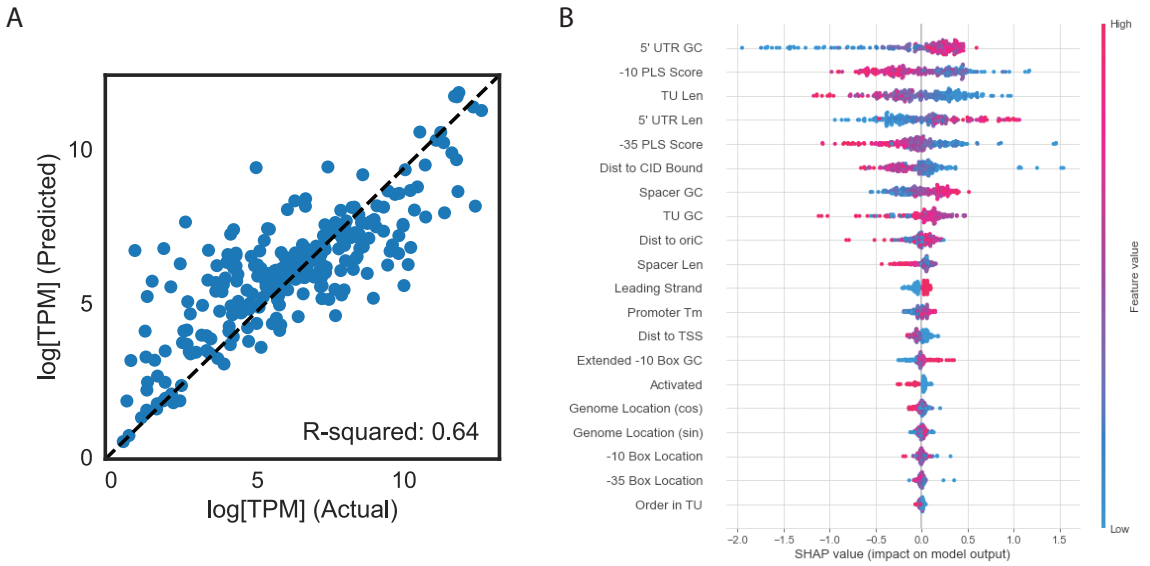


Figure 2.7: Machine learning model of expression. **(A)** Parity plot of best-performing random forest model, comparing predicted and actual transcript levels for 254 genes in a lockbox dataset. **(B)** Summary plot of feature importance calculations, generated with the *shap* package [55] Each gene from the lockbox set is represented by a point in each row of the plot. The features are ordered top-to-bottom based on their average absolute impact on model output (importance). The 0 point on the x-axis represents a baseline output of the model which is estimated by the *shap* package statistically.

during initial exploration of features. In particular, longer, GC-rich 5' UTRs were associated with higher expression by the model. Thus, overall, our Bitome-based machine learning model of *in vivo* expression in E. coli achieved good performance while highlighting both expected and unexpected sequence-based features contributing to transcript levels.

2.3 Discussion

Overall, the Bitome exhibits several important characteristics. Firstly, it unveils the uneven distribution of genomic features and positions. Secondly, it accurately captures the density of features in both coding and intergenic regions at high resolution. Thirdly, it reveals a higher occurrence of adaptive mutations at genomic positions with less feature density. Finally,

it facilitates the prediction of adaptively-mutated and essential genes, as well as quantitative prediction of transcript levels. Similar to the stoichiometric matrix, which is used to depict the reactome encoded in a genome in computational models [57], the Bitome serves as a binary, error-free knowledge-type object. While the stoichiometric matrix has been extensively utilized to study metabolic genotype-phenotype relationships [58, 59], the Bitome provides a comparable approach to understand the feature information encoded in a genome and supports predictive analysis based on that information.

Moreover, the Bitome has the potential for expansion to incorporate additional genomic features, allowing for the identification of further relationships beyond the core sequence-based features currently included. It can be applied to other genomes to examine and characterize their feature information distributions. Creating Bitomes for different strains would enable comparative analysis of feature information. Machine learning techniques like generative adversarial networks could be trained on a series of Bitomes from various strains to uncover underlying principles of genome organization that may not be evident in a single Bitome. These principles could serve as a foundation for the design of novel genomes. Additionally, analyzing Bitomes at the gene cluster level could enhance the prediction of gene function across species through synteny analysis. Overall, the Bitome serves as an organized, systematic representation of genomic information, offering a platform to unravel the "meaning" embedded within genomic sequences.

2.4 Methods

Assembling genome features

The E. coli strain K-12 substrain MG1655 reference genome (Reference Sequence NC_000913.3) was downloaded from NCBI in GenBank format. The reference was parsed using the SeqIO.read function from Biopython [60] (version 1.74). This reference genome defines the genomic positions. The following genomic features and their genomic locations were parsed from the reference genome: coding genes (CDS), pseudogenes, RNA-coding genes, insertion elements, repeat regions, and the origin of replication. Clusters of orthologous groups (COGs) functional annotations for genes from the reference genome were downloaded from NCBI [61] and linked via locus tag (b-number). Protein features were obtained from the GEM-PRO pipeline in the ssbio Python library [62] and linked to CDS from the reference genome by locus tag. Regulatory features were downloaded from RegulonDB [42] (version 10.0). The following regulatory features were parsed from RegulonDB: operons, transcription units, promoters (including -10 elements, -35 elements, and transcription start sites [TSS]), transcriptional and translational terminators, transcriptional and translational attenuators, Shine-Dalgarno sequences, riboswitches, transcription factor binding sites, and regulons (including sigmulons). Promoters not linked to a transcription unit were excluded. Genes from the reference genome were linked to operons and transcription units from RegulonDB via the locus tag. RegulonDB operons and transcription units not linked to a gene from the reference genome were excluded, and vice-versa. Independently-regulated gene modules [45] identified via independent component analysis (ICA) were linked to reference genome genes by locus tag.

Constructing the Bitome

Genome features were assembled into a sparse matrix using SciPy's [63] sparse matrix package. Each row represents a different genomic feature, and each column corresponds to a genomic position. Each element at row i and column j in the matrix has a value of either 1 or 0; 1 indicates presence of feature i in column j, and 0 indicates absence. To preserve the binary nature of the matrix (only 1s and 0s), features with multiple types were split into multiple rows as appropriate. For example, the 64 codons and 21 amino acids (this genome includes selenocysteine) were each represented in their own set of rows. To avoid overlaps and loss of information, certain features were split into six rows. These rows corresponded to three 'frames' (calculated as mod-3 of the start location) for each of the two strands (forward and reverse). Features treated in this manner were: genes, codons, proteins, amino acids (and all amino acid-based structural information), COGs. Regulatory features were represented in two rows corresponding to the forward and reverse strands. Regulons, sigmulons, i-modulons and transcription factor binding sites were left as single rows as no strand-specific information is available.

Computing sequence coverages

The 'bit counts' associated with each genomic position were calculated by taking the column-wise sum of the assembled matrix. Sequence coverages for selected features were computed by extracting a sub-matrix with just the rows corresponding to the features in question, summing the resulting sub-matrix row-wise, and computing the count of non-zero elements in the resulting vector along the length of the genome. Bit densities (in bits per bp) for genes and other genomic features were calculated by extracting a sub-matrix corresponding to the genomic

range of the feature in question, computing the sum of that sub-matrix, and dividing by the length of the genomic range.

Assembling and mapping ALE mutations

ALE mutations were downloaded from ALEdb [52] (version 1.0). SNPs based on reference sequence NC_000913.3 were selected. SNP density by genomic feature was calculated by determining the genomic positions with a 1 annotated for said feature (as described above) and dividing the total sequence length for that feature into the number of SNPs located at any of the feature's locations.

Computing mRNA secondary structure

mRNA minimum free energy structures were calculated with Nupack [64] in sliding 100 bp windows across the reference genome. A genome-wide average G was calculated; 'tight' regions were defined as those with minimum free energies in the lowest 10%, genome-wide.

Classifying genes with ALE SNPs

The scikit-learn (version 0.22.2) machine learning package was used to predict coding genes with ALE SNPs [65]. For each of 4186 coding genes, the Bitome matrix region corresponding to that gene's location was extracted. Each gene matrix was summed column-wise to create a gene feature vector. These feature vectors were transposed and concatenated into a gene feature matrix with dimensions 4186 (coding genes) $\times$ 1634 (Bitome features). The gene feature matrix was min/max normalized. A target label vector was generated by checking the location range of each gene for a SNP in ALEdb; if at least one was found, a 1 was placed in the target label vector; 0 otherwise. There were 2923 coding genes observed with SNPs, and 1263 without. 20%

of the data (evenly-weighted by class) was held out to generate a lockbox test dataset for final model evaluation.

The training data (gene feature matrix without lockbox data) still had a roughly 2-to-1 class imbalance. Thus, the majority class (SNP) was randomly down-sampled for all model training and cross-validation discussed below. Different classification models were evaluated for their performance on the training data. Adaptive boost, logistic regression, support vector machine, and random forest classifiers from scikit-learn - along with the XGBoost classifier from XGBoost version 1.0.2 [66] and an artificial neural network implemented with Tensorflow Keras - were run through 5-fold cross validation with five different downsampled training sets (Fig. A.2A). This same cross validation was performed after shuffling target labels as a negative control to obtain the expected accuracy of 50% (guessing), and with only the nucleobase features included. Hyperparameters for all models were optimized using a 5-fold randomized search cross validation approach.

Final model performances were assessed by re-training each hyperoptimized model on five downsampled versions of the lockbox test set. Based on this assessment, a support vector machine with the following non-default parameters was selected as the final model: penalty='l1', dual=False, C=0.1. Model coefficients for assessing feature importance were accessed using the *coef_* attribute.

Classifying essential genes

Essential gene labels were obtained from the Keio collection [53]. The scikit-learn package was again used for the classification workflow. Train and test sets were defined the same way as for ALE SNPs, except that mean instead of sum was used to collapse each gene sub-matrix into

a feature vector. There were 294 essential genes (class 1) and 3892 non-essential genes (class 0).

The same classifiers used for predicting ALE SNPs were tested for classifying essential genes. To address the large class imbalance, class frequency-weighted loss functions were used (for example, using the class_weight='balanced' argument for the scikit-learn classifiers). Models were initially assessed using 5-fold cross validation. Hyperparameters were optimized as with ALE SNPs.

Final performances were assessed by re-training each hyperoptimized model on the full training set and predicting based on the lockbox test set. Based on this assessment, a support vector machine with the following non-default parameters was selected as the final model: penalty='l1', dual=False, C=0.1, class_weight='balanced'. Model coefficients for assessing feature importance were accessed using the coef_ attribute.

Acknowledgements

This research was supported by the Novo Nordisk Fonden (NNF10CC1016517).

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

- **CR Lamoureux**, KS Choudhary, ZA King, TE Sandberg, Y Gao, AV Sastry, PV Phaneuf, D Choe, BK Cho, and BO Palsson. 2020. "The Bitome: digitized genomic features reveal fundamental genome organization." *Nucleic acids research*, 48(8):10157-10163. The dissertation author was the primary author.

Chapter 3

Escherichia coli functional non-coding regions are highly conserved

Rapid accumulation of microbial genome sequences enables large-scale studies of sequence variation. Most existing studies focus on coding regions to study amino acid substitution patterns in proteins. However, non-coding regulatory regions also distinctly influence physiologic responses. To assess intergenic sequence variation, we identified non-coding regulatory region alleles across 2,350 Escherichia coli strains. This "alleleome" consists of 117,781 unique alleles for 1,169 reference regulatory regions (transcribing 1,975 genes) at single base-pair resolution. We find that non-coding sequences are overall quite conserved; 64% of nucleotide positions are invariant, and variant positions vary in just 0.6% of strains on median. Non-coding alleles are sufficient to recover *E. coli* phylogroups. Critically, we find that functional non-coding regions

such as core promoter elements, transcription factor binding sites and transcription start sites are significantly conserved compared to un-annotated regions, especially when located upstream of essential or highly-expressed genes. However, variability in conservation of transcription factor binding sites is significant both within and across regulons. Finally, we contrast mutations acquired during adaptive laboratory evolution with wild-type variation, finding that the former preferentially alter positions that the latter conserves. Overall, this analysis highlights the wealth of information found in *E. coli* non-coding sequence variation and expands pangenomic studies to non-coding regions at single-nucleotide scale.

3.1 Background

Rapidly falling costs have yielded an explosion in complete genome sequences across organisms. Far from the first microbial genome assemblies almost 30 years ago, this wealth of sequence data necessitates genetic analyses that span thousands of genomes simultaneously [67]. Pathogen genomes are particularly well-represented due to sequencing surveillance efforts [68,69]; understanding genotype-phenotype relationships for these bacteria is critical. Pangenome analysis serves as a key tool towards this end [32]. Pangenome analyses have defined core (conserved) and accessory (variable) genomes for major microbial species [33] as well as identifying pangenome openness, or the continued discovery of unique genes as more genomes are sequenced [70]. Advancements in understanding antimicrobial resistance [34], virulence [36], and metabolism [35] have all been empowered by pangenomics.

However, pangenome analysis has been primarily focused on coding regions. While coding sequences make up the majority of a typical microbial genome [37], non-coding regions play an outsize role in manifesting phenotype from genotype. In particular, promoters [51, 71, 72] and

5' untranslated regions (5' UTRs) [73, 74] play a key role in executing the central dogma by modulating transcription and translation of operons. The core promoter features driving RNA polymerase binding via sigma factor recognition - the -10 and -35 boxes - are well known [24–27]. In turn, transcription start sites (TSS) have also been systematically identified [23] at base-pair resolution. Transcription factors (TFs) influence transcription in response to environmental stimuli via specific recognition of transcription factor binding sites (TFBS) within promoters and 5' UTRs [19, 28–30]. Transcriptional attenuators also play an important role in regulating expression of certain genes [31].

All of these sequence features are encoded differently from genes. They are therefore primarily located in non-coding regions, rendering them invisible to pangenome analyses focused solely on coding genes. Quantifying variation and conservation within these regions would shed light on the evolutionary pressures affecting control of expression. Because of the fine-grained nature of these critical sequence features, a base-pair resolution view of non-coding variation amongst wild-type genomes is warranted. Coding sequence pangenome analyses typically focus on presence/absence of homologous gene clusters within each strain or organism studied. To analyze intergenic sequence variation, a non-coding alleleome must be established, where "alleleome" refers to the aggregation of alleles for all sequence-based features of interest across a species [75–77]. Such a construct would serve to more explicitly link pangenomics to the deep literature surrounding molecular evolution and variation, which also focuses primarily on coding regions [78–83].

We therefore built a non-coding alleleome for Escherichia coli, focusing on promoter and 5' UTR regions. We amassed 2,350 fully-sequenced *E. coli* strains from across the phylogenetic tree, isolated from a variety of hosts. From the reference strain K-12 MG1655, we extracted 1,169

well-annotated non-coding regions that regulate transcription of 1,975 genes. We then identified and aligned alleles for these intergenic regions across the 2,350 strains. The resulting alleleome contains 117,781 unique alleles comprising over 400,000 base positions. Overall, we find the *E. coli* non-coding alleleome to be remarkably conserved. Furthermore, we: 1) cluster strains based solely on non-coding alleles, recovering phylogroups; 2) quantify variation and conservation within key sequence features; 3) identify essentiality and high expression as drivers of feature-specific conservation; 4) characterize variation in conservation across transcription factor binding sites; and 5) contrast functionally-impactful adaptive laboratory evolution mutations with wild-type variants. Taken together, the *E. coli* non-coding alleleome and analyses enabled by it represent an important expansion of large-scale genome sequence analysis to less-studied regions.

3.2 Results

3.2.1 The *E. coli* non-coding alleleome captures variation across a broad range of strains and regulatory features

In order to construct the *E. coli* non-coding alleleome, we identified all nucleotide sequence variants (alleles) for each of 1,169 well-annotated non-coding regions (from the reference strain K-12 MG1655) across 2,350 complete-genome, wild-type (WT) *E. coli* strains (Fig. 3.11A). We first amassed 2,350 completely-sequenced *E. coli* strains from BV-BRC [14]. The strains represented 14 distinct phylogroups as defined by ClermonTyping [86]. The majority belonged to *E. coli* sensu stricto groups, while 109 strains come from more distantly related clades or the fergusonii and albertii groups (Fig. 3.1B). The majority of strains with known hosts (67%) were isolated from humans, although other common domestic animals also provided strains (Fig. 3.1C). Bodily

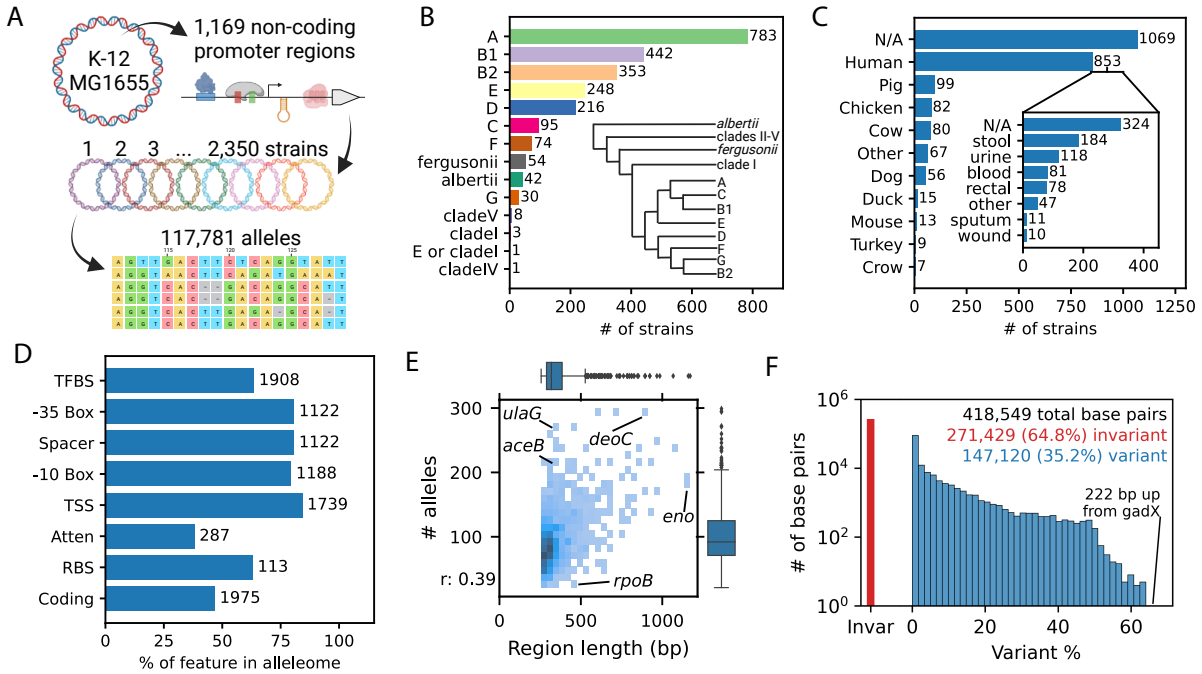


Figure 3.1: Constructing the *E. coli* non-coding alleleome. **A)** Schematic representation of *E. coli* non-coding alleleome construction. 1,169 non-coding promoter/5' UTR regions from reference strain *E. coli* K-12 MG1655 were mapped across 2,350 pangenome strains using BLAST [84]. The resulting 117,781 alleles across all regions were aligned within each region using MUSCLE [85] to create the *E. coli* non-coding alleleome. **B)** Strain counts for each of 14 phylogroups assigned by ClermonTyping [86] (phylogenetic tree adapted from ClermonTyping publication). Note: phylogenetic tree is not to scale. **C)** Breakdown of *E. coli* strains by host common name. Inset indicates bodily fluid/tissue of origin for strains isolated from human hosts. **D)** Counts and percentages of non-coding features from model strain K-12 MG1655 included in the non-coding alleleome. TFBS = transcription factor binding site, TSS = transcription start site, Atten = transcriptional attenuator, RBS = ribosome binding site. **E)** 2-D histogram comparing length of aligned non-coding regions (n=1,169) to number of distinct alleles found for that region across the alleleome. r = Spearman's r. Note: the minimum possible allele length for a region is 250 bp; 50 bp downstream from gene start + 200 bp upstream from TSS if TSS is at gene start (i.e. no 5' UTR). **F)** Histogram of variant percentages (i.e. percentage of non-dominant base pair) at each distinct aligned position in the *E. coli* alleleome. Blue histogram indicates variant % distribution for positions with non-zero variation. Red bar indicates the number of invariant base pairs.

excretions were the most common known sources of human-isolated strains.

These key non-coding regions capture 35.7% of the total non-coding positions in the reference strain and control the transcription of 1,975 genes (Fig. 3.1D). Importantly, these regions included majorities of key non-coding features, including: 84% of TSS, 80% of core promoters,

and 64% of TRBS. We then searched for these reference non-coding regions across the full set of *E. coli* strains, extracting homologous sequences from the expected local regions upstream of homologous genes in these other strains. Then, for each set of sequences corresponding to a reference region, we used multiple sequence alignment to determine the WT occurrence of every nucleotide (including indels) at every position.

In total, we identified 117,781 distinct alleles across all regions and strains; these alleles comprise the *E. coli* non-coding alleleome. The median length of an aligned non-coding region was 319 base pairs, and the median region had 92 distinct alleles (Fig. 3.1E). Region length and number of alleles were weakly correlated (0.39, Spearman). The promoter and 5' UTR of *deoC* - encoding pyrimidine catabolism enzyme deoxyribose-phosphate aldolase - contained notable variation, with 294 distinct alleles in the 896-bp region. The upstream region of *eno* (encoding glycolysis and degradosome enzyme enolase), despite being the longest region considered at 1,171 bp, had just 184 unique alleles. Much of *eno*'s upstream region overlaps with the upstream *pyrG* gene, which may influence conservation in this multi-purpose region. RNA polymerase core subunit gene *rpoB*'s region has just 32 distinct alleles despite a length of 442 bp, highlighting a level of conservation commensurate with this gene's essential role. Ascorbate degradation gene *ulaG* and glyoxylate cycle enzyme *aceB* featured particularly variable regulatory regions given their relatively short lengths. Overall, we assembled 418,549 aligned base pairs; of these, 65% are completely invariant (Fig. 3.1F). The median variant percentage for variant positions was just 0.6%, highlighting an overall substantial level of sequence conservation across the non-coding alleleome. However, specific base pairs are particularly variant. The most variant base pair in the alleleome is found 222 base pairs upstream of the *gadX* gene start; an indel results in the dominant "base" being a gap found in 34.4% of genomes, with A and G in 34.0% and 31.5% of

cases, respectively.

3.2.2 Non-coding alleles recover phylogroups and highlight outliers

We next investigated the co-occurrence of non-coding alleles across strains. Hierarchical clustering of strains - with similarity defined as the fraction of shared alleles across the 1,169 regions under consideration - yields 14 clusters, matching the number of phylogroups (Fig. 3.2A). Overall, strains within the same phylogroup tend to be assigned to the same cluster (Fig. 3.2B). Thus, as expected, non-coding alleles are more shared within phylogroups, and indeed are sufficient to discriminate between phylogroups in most cases. Interestingly, A, B1, and C - while most similar within their respective groups - are nonetheless similar enough with each other to be grouped together in Cluster 2. One strain identified by the phylogenetic method as ambiguous between phylogroup E and cladeI clusters with all phylogroup E strains, again highlighting that non-coding alleles alone carry sufficient information to determine phylogroups.

Interestingly, clusters 4 and 5 contain single outlier strains from phylogroups B1 and A, respectively. A second B1 outlier strain appears in cluster 8 with most of the phylogroup D strains. Closer examination of the median pairwise distances of these two strains with all other B1 strains confirms that these strains indeed do share much fewer alleles than a typical pair of B1 strains (Fig. 3.2C). We then further inspected B1 strain GF4-3 (the cluster 4 outlier) by identifying, for each of the non-coding regions, whether this strain's allele was completely unique within phylogroup B1 or shared with at least one other B1 strain. By analyzing the clusters of orthologous groups (COG) distributions of genes transcribed from the regions within these unique and shared groups, we identified the particular functional characteristics that contribute disproportionately to this strain's distinctiveness (Fig. 3.2D). In particular, this strain has nearly

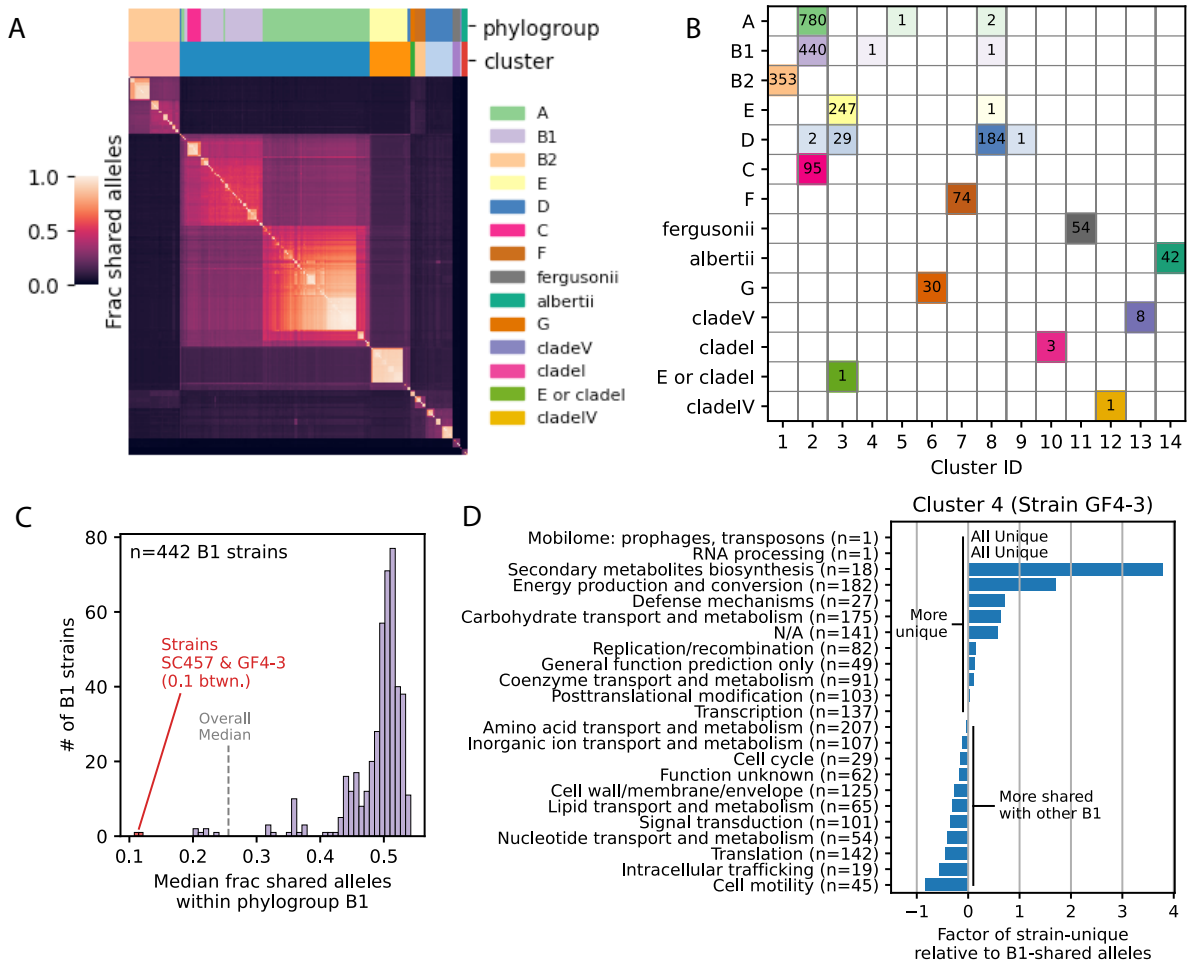


Figure 3.2: Non-coding allele clusters capture *E. coli* phylogroups. **A)** Clustermap of 2,350 *E. coli* strains, based on non-coding alleles. Heatmap displays a similarity matrix based on the fraction of shared alleles across 1,169 distinct non-coding regions (i.e. a value of 1.0 indicates that strains have identical sequences at all 1,169 non-coding regions. Upper colorbars indicate phylogroup (determined separately using ClermonTyping [86]; see Materials and Methods) and cluster assignment based on hierarchical clustering of distance matrix (1 - similarity matrix). Legend colors correspond to colors from the top colorbar row. **B)** Cluster vs. phylogroup assignment matrix. Rows are phylogroups, columns are clusters; e.g. all 30 strains in phylogroup G are in cluster 6, and cluster 6 contains no other strains. Phylogroup color scheme same as panel A. n=2,350 total strains. **C)** Histogram of pairwise similarity (defined as fraction of shared alleles) within phylogroup B1. Overall median = median pairwise similarity across all strain comparisons (i.e. median of all entries in heatmap from panel A. Strains GF4-3 and SC457 are outlier B1 strains, found in clusters 4 and 8 respectively. **D)** Relative enrichment of fractions of clusters of orthologous groups (COGs) for genes transcribed by promoter alleles found uniquely in strain GF4-3 (cluster 4) vs. by promoter alleles shared with at least one other B1 strain. For each COG, value is: (fraction of unique allele-transcribed gene COGs) / (fraction of shared allele-transcribed gene COGs) - 1; i.e. a value of 0 indicates that the COG comprises an equal fraction of the unique and shared sets. E.g., the "Defense mechanisms" COG is about 1.6x more represented in the unique set than the shared.

four times more unique alleles related to secondary metabolite biosynthesis and nearly two times more related to energy metabolism. These unique characteristics may stem from this strain's host, the guineafowl (the only strain in this dataset isolated from this African bird).

3.2.3 *aceB* intergenic region provides a case study for analysis of sequence variation within functional sites

As a case study, we focused on a specific 331-bp non-coding region - the 5' UTR and promoter region upstream of *aceB* (malate synthase A; a key enzyme in the glyoxylate cycle). We selected this region due to its relatively large number of alleles despite its short length. This region was identified in 2,340 of 2,350 strains, with the most common allele appearing in 20.2% (473/2,340) of strains (Fig. 3.3A). While some common alleles dominate, a variety of more niche alleles are also present. 90% of strains are accounted for by just 22% of the alleles; however, accounting for 99% of strains requires 90% of alleles. This region contains one transcription start site (TSS) with -10 and -35 elements, 11 TF binding sites of 4 distinct TFs, a transcriptional attenuator, and the very end of the next upstream gene, *metA* (homoserine O-succinyltransferase; catalyzes first step in methionine biosynthesis) (Fig. 3.3B). 20 significantly variant positions (those with variant base pairs present in at least 15% of strains) are mostly found upstream of the core promoter region, with a particular concentration in a specific IclR binding site. An additional 52% (173/331) of positions have minor variants, and 42% (141/331) are invariant.

Assessing variant presence in different genomic features provided a more detailed view of variation in this region. For example, while 33% of base pairs in this region are annotated as being part of at least one TF binding site, only 28% of all variant base pairs are found in TF binding sites (a factor of 0.15 fewer) (Fig. B.1). Conversely, positions with no annotation accounted for

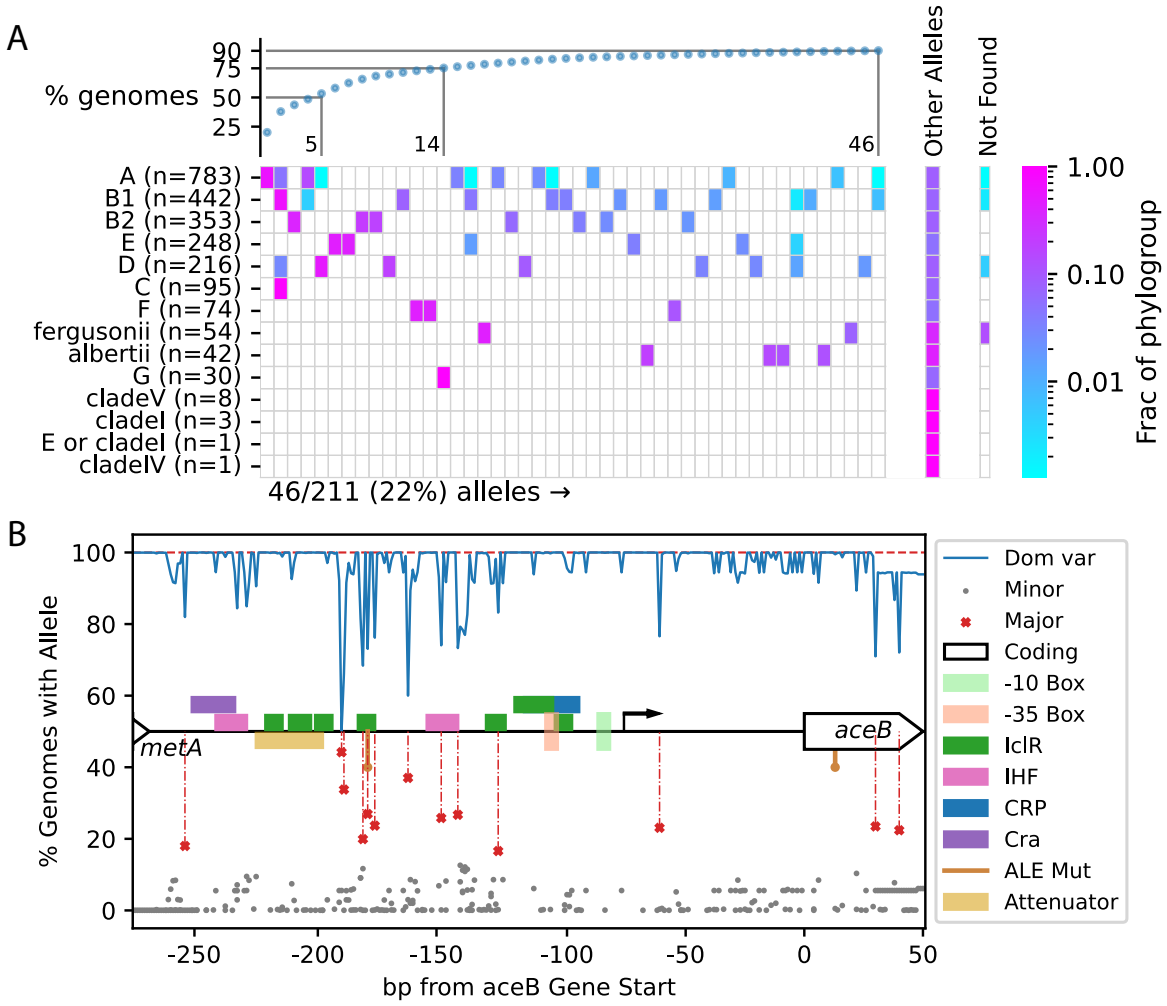


Figure 3.3: Alleleome for a single intergenic region; a case study. Statistics for a single region from the non-coding alleleome, comprising the 5' UTR and promoter region for *aceB* (malate synthase A). **A)** Heatmap of phylogroups vs. common alleles for this region. Colorbar is scaled per row (phylogroup); e.g. hot pink in phylogroup row G, allele column 14 indicates all G strains (fraction 1.0) have this allele. Alleles are sorted left-to-right in decreasing order of fraction of strains with allele. Scatterplot above heatmap indicates cumulative fraction of strains covered by corresponding number of alleles. Not found indicates fraction of each phylogroup for which this region was not found at all (no allele). **B)** Depiction of sequence features from reference strain (K-12 MG1655) in this 331-bp non-coding region (central track), along with dominant (blue line) and variant (gray dots and red crosses) positions by percentage of genomes found in. Major variants defined as those present in at least 15% of strains.

42% of the sequence but 49% of the variant base pairs. The core promoter elements - TSS, -10 and -35 elements - are particularly lacking in variants relative to their sequence exposures. No

transcriptional terminators or ribosome binding sites are annotated for this region. While these observations hint at potential conservation patterns, one example non-coding region is insufficient to quantify systematic WT variation trends.

3.2.4 Aggregating non-coding alleles across the genome reveals conservation in functionally important regions

We repeated the *aceB* analysis across all 1,169 non-coding regions and combined the results, revealing genome-wide trends in conservation within the non-coding alleleome (Fig. 3.4A). On median, 16% of a region's most present alleles capture the sequence diversity of 90% of genomes, while a median of 76% of alleles are required to span 99% of genomes (Fig. 3.4B). However, these distributions are quite broad - indeed, some regions are highly conserved, needing as few as 23% of alleles to cover 99% of genomes (region upstream of ribosomal protein *rpsM*). The most common allele covers a median of 33% of genomes; however, certain highly conserved regions - again including the region upstream of ribosomal protein *rpsM* - are covered almost entirely by a single dominant allele (Fig. 3.4C). On the median, these regions are 65% invariant base pairs, 32% minor and 2% major variants (Fig. 3.4D). There is notable variability across regions: for example, 31% of base pairs upstream of gluconeogenesis gene *pck* (encoding phosphoenolpyruvate carboxykinase) have major variation.

Most importantly, combining observations of variation across non-coding regions allows for assessment of conservation within annotated features. On median, non-coding base pairs without annotation vary just 3% more than expected based on their sequence coverage, indicating minimal deviation from the background mutation rate. All non-coding features aside from attenuators vary significantly less than unannotated regions (Mann-Whitney U, FDR¡0.01). Ribosome

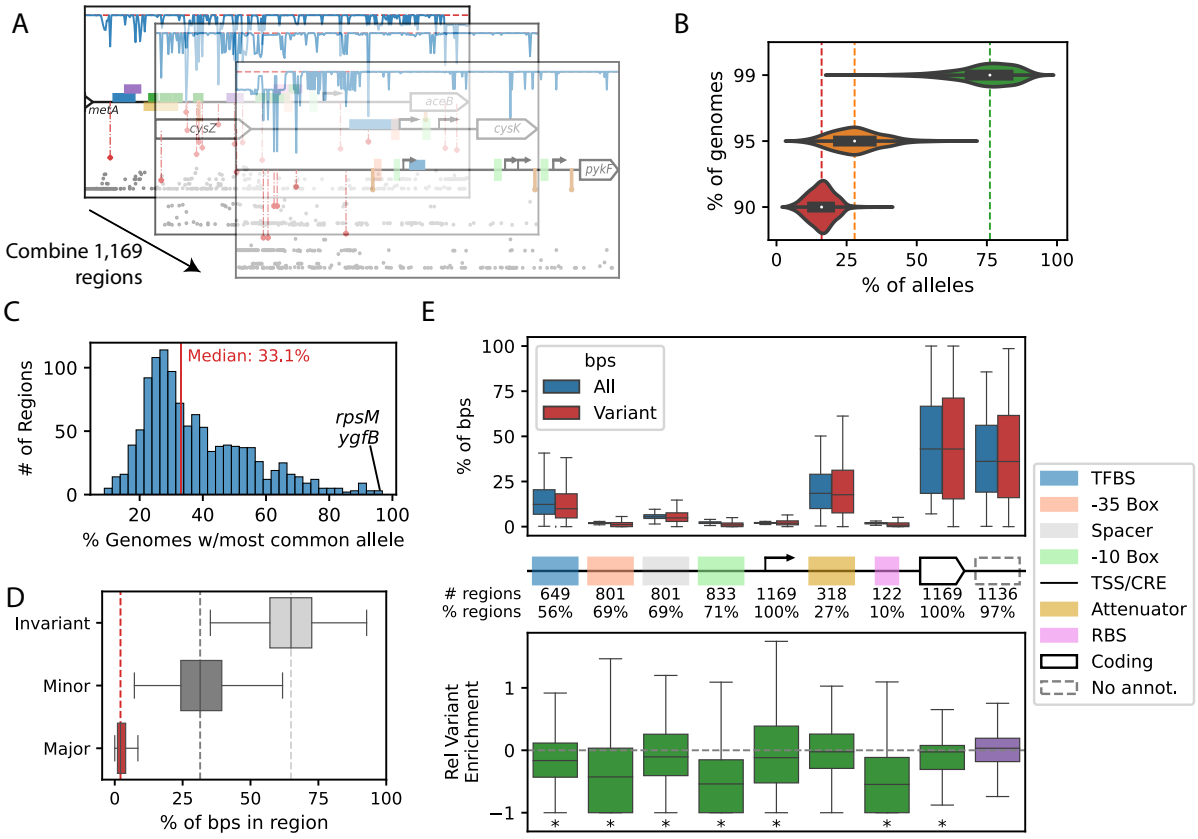


Figure 3.4: Summary statistics of sequence variation for all 1,169 non-coding regions' alleles. **A)** All summary statistics for each of 1,169 non-coding regions (represented by dashboards from Fig. 3.3 were aggregated to investigate whole alleleome properties. **A)** Violin plot showing distributions of allele percentages needed to cover 90/95/99% of genomes for a given non-coding region. **A)** Histogram of % of strains covered by the most common allele; e.g. the non-coding regions upstream of *rpsM* and *ygfB* are identical in over 95% of strains in which the region was found. **A)** Box plot showing distributions of variant types across non-coding regions. **A)** Distributions of sequence and variant base pair percentages across annotation categories (see Figure 3C). #/% regions = number/percentage of regions that have at least one base pair annotated with the indicated category (e.g. 649/1,169 (56%) of non-coding regions have at least one TF binding site). Asterisks = significant difference with no annotation regions (U test, FDR 0.01).

binding sites and the core promoter elements (-10 and -35) are the most conserved sequences in non-coding regions. RBS, -10 elements, and -35 elements are on median 58%, 57%, and 46% less variant than base pairs without functional annotation, respectively (Fig. 3.4E). TF binding sites, the spacer between the core promoter elements, and the TSS and core recognition element

(CRE) are all more conserved than unannotated regions (20%, 14%, 15% respectively). Coding regions included in this alleleome due to opposite strand overlap are just 6% less variant than unannotated non-coding base pairs, and only 3% less variant than expected based on sequence coverage. However, overlap with coding regions does significantly reduce variation in TF binding sites, spacers, CREs and attenuators (Fig. B.2A). This effect is minimal when considering non-coding features that straddle coding region boundaries, suggesting that variation within any given feature is selected relatively uniformly (Fig. B.2B).

Additionally, conservation in upstream non-coding regions relates to the functions of their gene products. Regions expressing genes in clusters of orthologous groups (COG) categories such as "Translation, ribosomal structure and biogenesis" and "Replication, recombination and repair" are amongst the most conserved (Fig. B.3). Conversely, metabolic COGs appear to support more non-coding variation, such as amino acid metabolism and secondary metabolite biosynthesis. Non-coding regions transcribing at least one essential gene are significantly more conserved than those that do not transcribe any essential genes (Fig. B.4A). These essential-transcribing regions also have significantly more conserved transcription factor binding sites and promoters (Fig. B.4B). The effect size is largest for the -10 and -35 elements of the promoter, highlighting the expected significance of these sigma factor binding regions. Similarly, non-coding regions also differ significantly in conservation depending on their baseline expression level (Fig. B.4C). As with essential-transcribing regions, this conservation is also prevalent in the most critical promoter regions (Fig. B.4D).

3.2.5 Transcription factor binding sites exhibit significant variation in conservation

The non-coding alleleome enables a detailed investigation of conservation within transcription factor binding sites. We identified 22 major transcription factors that have at least 10 binding sites and whose activity explains notable variation within the PRECISE-1K expression compendium. The median percentage of invariant base pairs within the binding sites of these TFs is not significantly correlated with the percentage of expression variation explained (Fig. 3.5A). Most of these transcription factors have binding sites with a wide range of conservation (Fig. 3.5B). Central carbon metabolism regulator Cra's binding sites are the most conserved, with a median of 84% invariance. Nucleotide metabolism regulator PurR's binding sites are consistently conserved; only one site falls below 70% conservation. A subset of these TFs are further identifiable as dual regulators, with at least 10 binding sites annotated for both activation and repression roles. Mostly, this distinction does not result in a difference in binding site conservation (Fig. 3.5C). However, the TF and nucleoid-associated protein IHF has significantly more conserved repressor sites than activator. IHF is known to be able to bind to DNA in a non-specific manner and may even be redundant with AT-rich upstream regions in some cases [87]. Example CRP binding sites highlight the range of conservations observed within binding site sequences (Fig. 3.5D). A binding site upstream of *cyoA* has no completely conserved base pairs, while a CRP binding site regulating *rpoH* expression has only one position with any variation. Interestingly, the binding site upstream of *cyoA* exhibits particular variation across the alleleome in a relatively high-information region of the reference strain CRP motif, possibly indicating a functional impact of these variants on CRP activity.

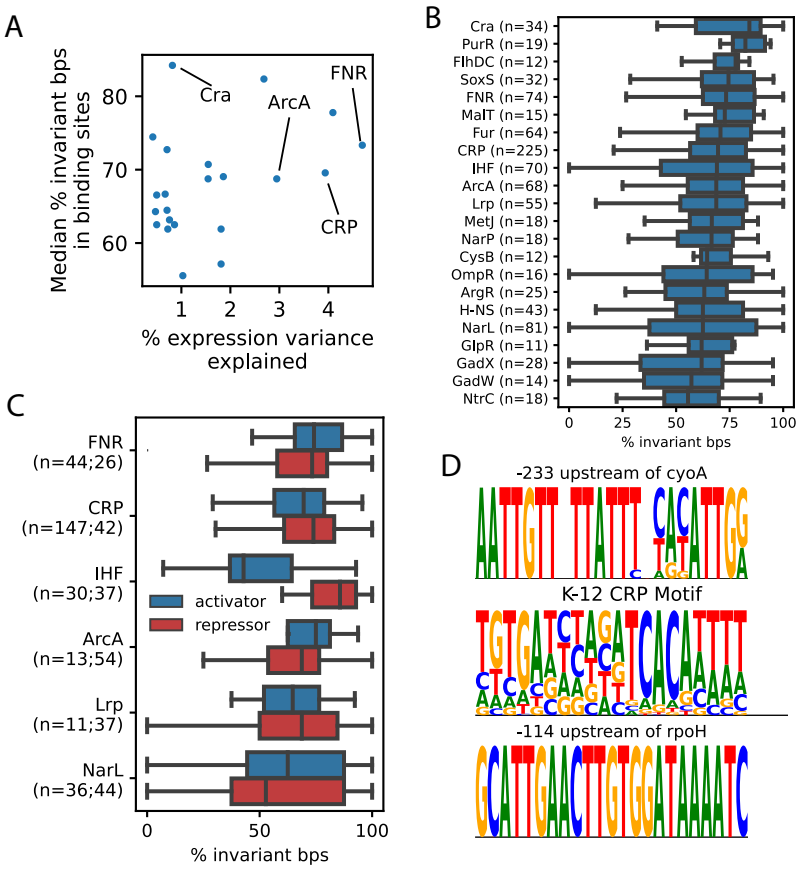


Figure 3.5: Transcription factor binding sites exhibit a wide range of conservation. **A)** Comparison of expression variance explained to median percentage of invariant base pairs within binding sites for 22 major transcription factors (min 10 binding sites). Explained variance % is computed based on percentage of expression variance in PRECISE-1K expression compendium [54] explained by TF's iModulon (a gene grouping capturing the independent effect of the regulator). **B)** Distributions of invariant base pair % for binding sites of 22 major transcription factors. **C)** Distributions of invariant base pair % for major regulators with dual regulatory effect (min 10 annotated binding sites for each mode of regulation; individual sites annotated as "dual" removed; effect data from RegulonDB [19] **D)** Example sequence logos for poorly conserved (top) and highly conserved (bottom) CRP binding sites. CRP motif from reference strain K-12 MG1655 (from RegulonDB) is shown in middle. Note: all base pairs in *cyoA* site have variants; some variants are too small to visualize. Note: position 1 in *rpoH* site has 7/2350 variants.

3.2.6 Laboratory adaptive mutations are more likely than natural variants to impact functionally-relevant features

In contrast with natural variants, adaptive laboratory evolution (ALE) exerts selective pressure for cells to adapt to a particular stress or growth mode. ALE's preference for high-impact

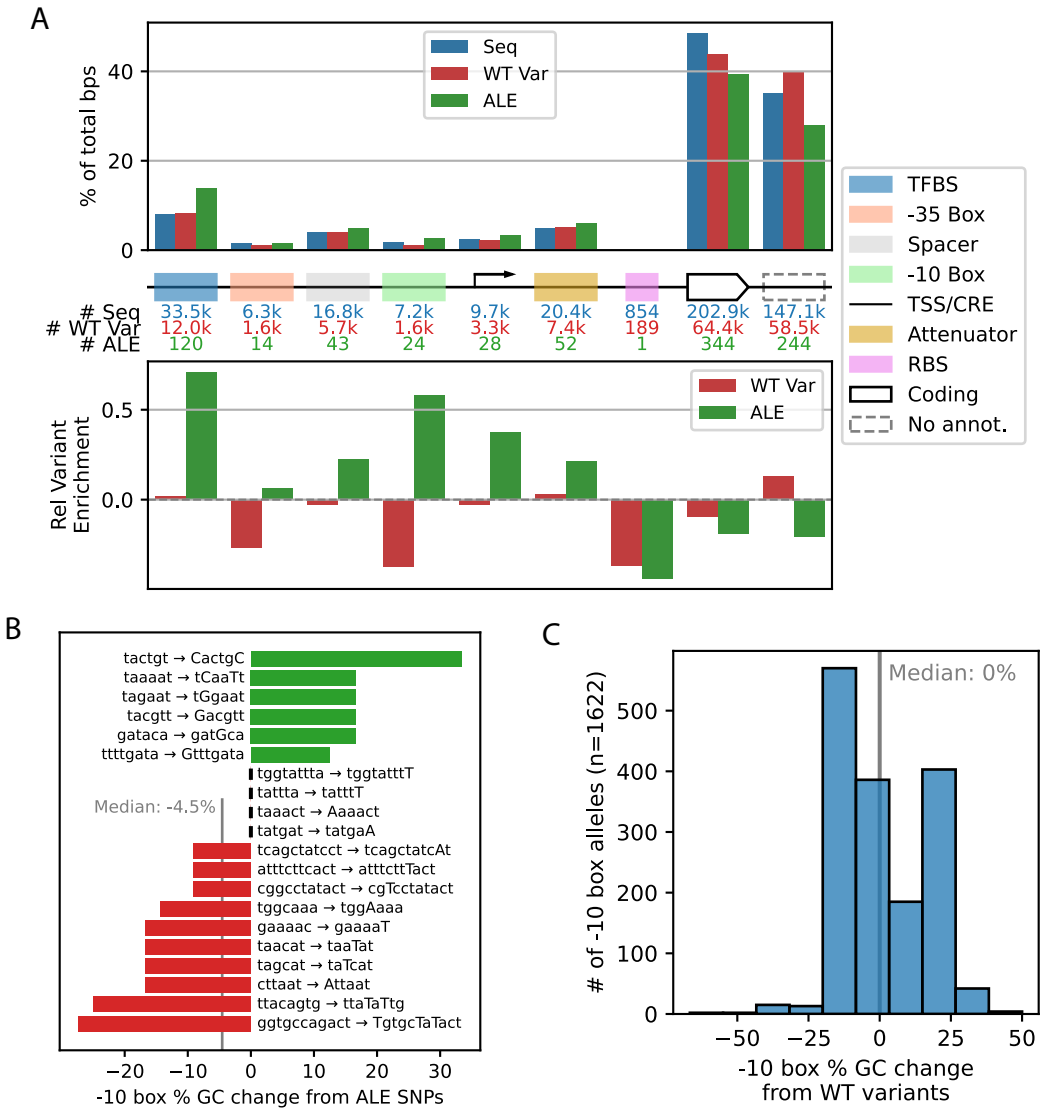


Figure 3.6: Adaptive laboratory evolution (ALE) mutations are over-represented in wild type-conserved non-coding regions. **A)** Breakdown of percentages of all non-coding base pairs considered by annotation category. In upper panel, blue bars represent % of all base pairs annotated within each category; red bars represent % of all variant base pairs annotated within each category; green bars indicate percentage of all 1,174 non-coding ALE mutations present in each category. **B)** Effect of 20 ALE SNPs affecting -10 boxes on GC content. **C)** Effect of 1,622 distinct wild-type -10 box variant alleles on GC content relative to consensus -10 box sequence for that region.

mutations becomes clear when comparing the rates of ALE mutations in particular non-coding regions to wild-type variant rates. For example, ALE mutations are 75% more likely to occur

in TF binding sites than these base pairs' sequence exposure would suggest (Fig. 3.6A). Core promoter elements also exhibit this effect; -10 and -35 boxes are mutated 58% and 35% more often than expected. Not only are ALE mutations enriched in -10 boxes, but the mutations have a slight tendency to reduce the GC content of these regions (Fig. 3.6B). The -10 box is typically the upstream location of DNA strand unwinding for transcriptional bubble formation upon RNA polymerase binding; thus, decreased GC content is likely to increase transcription at these sites. Wild-type variants at -10 boxes don't tend to alter GC content on median (Fig. 3.6C).

3.3 Discussion

Here, we present a non-coding alleleome for Escherichia coli, providing a deep look at variation in critical transcriptional and translational control regions. We assemble 2,350 complete genomes across the E. coli phylogenetic tree and identify alleles for 1,169 reference non-coding regions across this set of strains. We cluster strains based on their non-coding alleles, finding these to be largely sufficient for distinguishing phylogroups. Centrally, we find that overall sequence variation in these non-coding regions is minimal; 64% of positions are completely conserved, and variation at the remaining positions is overwhelmingly minor. As hypothesized, core promoter features and binding sites are more conserved than non-functional positions. We also show that essentiality and high expression drive significant conservation, again concentrated in functionally critical promoter features. The alleleome also provides a rich understanding of conservation across transcription factor binding sites, highlighting significant variation both between and within different regulators' binding sequences. Finally, we contrast wild-type variation with mutations acquired during adaptive laboratory evolution, determining that adaptive mutations preferentially alter regions that natural variants conserve.

This analysis expands our understanding of natural sequence variation beyond coding regions. While 5' UTR and promoter regions constitute a minor fraction of the genome sequence, they encode the critical control functions that enable E. coli to adapt its transcriptome in response to environmental signals. Our understanding of precisely how these sequences influence expression in vivo - as opposed to via synthetic promoter libraries - remains limited. This non-coding alleleome provides a new dimension with which the function of these regions may be further elucidated. For example, models that aim to predict expression level directly from promoter sequence may benefit from understanding how conserved each base pair in the promoter region is; a similar approach is important for the function of AlphaFold(45).

The identification of two unique strains whose non-coding alleles do not cluster notably with any other phylogroups highlights a potential bias in complete E. coli genome sequences currently available. The E. coli strain GF4-3, isolated from a guineafowl, harbors distinct non-coding alleles from any other strain observed in this study. E. coli sequence diversity may be significantly more rich than we currently realize due to over-representation of strains isolated from a handful of host organisms.

E. coli's genome is largely dominated by coding genes; in the reference strain K-12 MG1655, 87% of base pairs are part of a coding gene [37]. As a result, many positions within the non-coding alleleome are technically also within coding regions. This situation may arise due to promoter regions found within operons, such that a promoter overlaps with the nearest upstream gene; strand differences, where a gene encoded on one strand is directly opposite a promoter region on the other; or divergent promoters, where a relatively small promoter region is shared between two genes transcribed in opposite directions. Our analysis indicates that, in general, coding positions vary at the expected rate based on their sequence coverage. However,

further study is needed to determine whether the coding or non-coding functions encoded in these regions are driving conservation patterns.

The non-coding alleleome's quantification of variation in transcription factor binding sites provides an opportunity for expansion of binding motif definition. Motifs aim to summarize the specific sequence required for binding of a transcription factor to DNA by combining the sequences of experimentally-determined binding sites and indicating the probability of finding each base at each position. Motifs are typically generated by combining binding sites controlling different transcription units within the same strain. Frequently, real observed binding site sequences differ significantly from a canonical motif. Thus, alleleome variation within the same binding site may provide an alternative information source for assessment of binding site sequence importance by allowing comparison of alternative sequences within a more similar sequence and functional context. Any time a new experimental binding site is identified, alleleome variation within the proposed site can be assessed to provide context for the likely strength or importance of the site. However, because we do also observe significant variation in conservation across binding sites, this approach may only provide one piece of information as part of a larger picture.

Overall, this E. coli non-coding alleleome quantifies base pair-level variation and conservation at genome- and species- scale. The data generated in this study provides a rich resource for analyzing non-coding regions in any E. coli genome. We believe that this type of analysis should be expanded to other organisms to enable comparative non-coding alleleomics. As sequence data continues to balloon, this study provides a blueprint for compiling, quantifying, and analyzing non-coding variation, revealing patterns of conservation and their relationship to phenotypic outcomes.

3.4 Methods

Assembling complete *E. coli* genome sequences

Complete *E. coli* genome sequences and metadata were downloaded from BV-BRC (formerly known as PATRIC) [14]. These genomes were subjected to the following quality control steps. Completeness and quality were verified by selecting genomes with "Contig L50" of 1 and "Contig N50" ¿ 4M. Furthermore, only genomes without ambiguous bases (i.e. only ACGT in sequence) were selected. Finally, genomes were selected only if they had coding sequences annotated (i.e. a GFF/FAA file was also downloaded). Phylogroups were assigned for each genome sequence using the ClermonTyping in silico tool [86]. Genomes annotated as "Non Escherichia" or "Unknown" were excluded. After these filtering steps, 2,350 complete genome sequences remained.

Generating coding sequence pangenome

A coding sequence pangenome was generated as described previously [88]. All FAA files for all amino acid sequences of all genes from all valid strains were combined into a single file and subjected to duplicate removal, yielding a listing of all 918,781 non-redundant protein sequences. This file was then provided to the CD-HIT protein sequence clustering program (v.4.8.1 [89] with the following non-default options: "-n 5 -c 0.8". This processing yielded 80,453 gene clusters. These clusters (and their constituent individual alleles) were then given unique identifiers and referenced back to the strain(s) from which they came.

Identifying reference non-coding regions and features

High-confidence transcription start sites (TSS) for the reference strain *Escherichia coli* K-12 MG1655 (genome accession number NC_000913.3, BV-BRC/PATRIC genome ID 511145.12) were accessed from RegulonDB [19]. This resource has been extensively manually curated and comes with additional annotation of non-coding and regulatory features for these high-confidence TSSes. 2,228 TSS were annotated as transcribing at least one coding gene. Each TSS was mapped to the first gene it transcribes using the Bitome [37]. Then, a sequence region starting from 200 base pairs upstream of the TSS through 50 base pairs downstream of the first gene's start codon was extracted for each of these TSS/first gene pairs. At this stage, a separate region was extracted for alternate TSSes transcribing the same first gene, even if the regions partially overlapped. These nucleotide sequences were then written to a FASTA file.

Searching for reference non-coding regions across all strains

For each pangenome strain, coding genes that appeared in a cluster with a K-12 MG1655 gene were selected. For each of these coding genes, the maximum upstream-from-gene-start length for a reference non-coding region was determined. A local search region spanning 100 base pairs further upstream from this maximum reference upstream length through 100 base pairs downstream of the pangenome strain's gene start was extracted. For example, if a reference non-coding region from K-12 MG1655 had a 150-bp 5' UTR, plus the additional standard 200 bp upstream from the TSS, the local search region in a pangenome strain for this non-coding region would start 450 base pairs upstream of the pangenome strain's gene that clustered with the reference strain gene transcribed by the reference non-coding region. Within each pangenome strain, all such search regions were combined into a single FASTA file and passed to create a

BLAST search database with the BLAST+ [84] program makeblastdb. Then, blastn was used to search for all reference non-coding regions against this strain- and region-specific database. For each pangenome strain, only BLAST matches for a reference non-coding region in the local search region upstream of the pangenome strain gene corresponding to the appropriate reference strain gene were kept. If multiple alignments were found within the correct local search region, the alignment with the lowest E-value was selected. For each match, the corresponding nucleotide sequence of the non-coding region allele from each strain was extracted from the strain's genome. Finally, all sequence matches for a given reference non-coding region were grouped together.

Building the non-coding alleleome

For each set of non-coding sequences corresponding to a particular reference non-coding region (non-coding alleles), the nucleotide sequences were aligned using multiple sequence alignment tool MUSCLE [85] with all default arguments. Aligned sequences with greater than 20% gaps were filtered out. Then, only non-coding regions with an allele found in at least 75% of strains were kept. At this point, due to alternate TSS for the same transcription unit, some non-coding regions could be subsets of others. Thus, for each set of alternate TSS, only the longest aligned set was selected for further analysis as the other regions would be subsets thereof. These steps led to the identification of 1,169 final regions that - with all of their alleles - comprise the *E. coli* non-coding alleleome.

Annotating alleleome base pairs with variant and feature information

For each aligned base pair in the alleleome, variant percentage was calculated as the percentage of strains that have the non-dominant base at that position. Then, using the Bit-

ome [37], each base pair was annotated for presence/absence of the following features: gene, TSS, core recognition element, -10 box, -35 box, -10/-35 spacer region, ribosome binding site (Shine-Dalgarno sequence), transcription factor binding site, and transcriptional attenuator. Furthermore, each non-coding region was annotated as essential or non-essential, with essential defined as any of the TSS in the non-coding region transcribing at least one gene annotated as essential in the Keio collection [53]. Each non-coding region was also assigned a baseline expression level category of Low, Medium, or High, based on the median of median expression levels across all genes transcribed from the region, using the PRECISE-1K definitions of the three categories [54]. Finally, clusters of orthologous groups (COG) categories were assigned to each non-coding region based on the unique set of COGs assigned to genes transcribed from each region (a non-coding region could be assigned multiple COGs).

Clustering strains by non-coding alleles

The linkage function from the SciPy [90] hierarchical clustering package was used on a pairwise distance matrix between all 2,350 strains, with non-default argument method='average'. The pairwise distance was constructed by taking the complement of a similarity matrix, where the similarity between two strains was defined as the fraction of all 1,169 non-coding regions for which the two strains had exactly the same allele. Then, flat clusters were computed using the Scipy fcluster function in 'maxclust' mode. The optimal 'maxclust' parameter was determined using a sensitivity analysis considering values from 2 through 25 inclusive and computing the mean silhouette score across all strains. This analysis selected 14 as the optimal number of clusters.

Acknowledgements

This research was supported by the Novo Nordisk Fonden (NNF20CC0035580).

Chapter 3 in part is a reprint of material submitted for publication in *Nucleic Acids Research Genomics and Bioinformatics*:

- **CR Lamoureux**, PV Phaneuf, BO Palsson, and DC Zielinski. 2023. "*Escherichia coli* functional non-coding regions are highly conserved." The dissertation author was the primary author.

Chapter 4

A multi-scale *Escherichia coli* expression and regulation knowledge base

Rapid accumulation of transcriptomic data necessitates scalable approaches to extract insights from this information. In this study, we constructed a comprehensive knowledge base for *Escherichia coli* K-12 MG1655, focusing on gene expression and regulation. The expression component comprises a high-quality RNA-seq compendium consisting of 1,035 samples, encompassing various growth conditions such as 9 different media, 39 supplements (including antibiotics), 42 heterologous proteins, and 76 gene knockouts. Utilizing this extensive resource, we uncovered global expression patterns and employed machine learning techniques to identify 201 modules that capture 86% of known regulatory interactions, forming the regulatory component. By leveraging these modules, we discovered two previously unknown regulons and quantified system-level

regulatory responses. Additionally, we integrated 1,675 curated, publicly-available samples into the knowledge base, enhancing its scope. Furthermore, we demonstrated workflows that enable the analysis of new data against this knowledge base, exemplified by deconstruction of regulation during aerobic transition. This resource not only sheds light on the *E. coli* transcriptome on a large scale but also presents a blueprint for top-down transcriptomic analysis of non-model organisms.

4.1 Background

Over the past decade, RNA sequencing (RNA-seq) has emerged as a powerful and efficient method to assess the expression state of cell populations. The availability of large RNA-seq datasets [45,91–94] has necessitated the development of big data analysis methods to enhance our understanding of transcription and regulation [45, 95–99]. As datasets continue to expand, these methods that can effectively convert this vast amount of data into biological insights increase in importance. Thus, a unified and comprehensive resource that integrates expression data, regulatory information, and big data analysis is desirable.

The analysis of large RNA-seq datasets from multiple sources can be complicated by batch effects that can hinder accurate analysis and interpretation. Consequently, mitigating these batch effects remains a crucial goal and an area of ongoing research [100, 101]. One possible strategy to mitigate this issue is the utilization of single-protocol, high-quality, and curated RNA-seq datasets. However, creating such datasets is time-consuming and costly.

Transcriptional regulatory networks (TRNs) are key tools for representing regulation within an organism. Constructing TRNs involves exhaustively characterizing the binding of regulators to promoter regions of target genes and their impact on gene transcription. Consequently,

inferring regulatory signals directly from an RNA-seq dataset, without prior knowledge of the TRN, would be a valuable component to a data-driven, top-down transcriptional resource.

Independent component analysis (ICA) [102] is a signal processing algorithm that outperforms other methods in extracting biologically relevant regulatory modules from gene expression data [103]. Application of ICA to publicly available prokaryotic expression data has consistently identified TRN modules across various organisms [45, 104–108]. ICA's effectiveness stems from its ability to identify independent sets of genes that exhibit consistent variation across samples, regardless of group size or overlapping membership. Hence, a dataset with sufficient scale and diversity in conditions is crucial for the successful application of this method.

In this study, we introduce an expression and regulation resource for *Escherichia coli* K-12 MG1655, a key model organism. The expression component, known as PRECISE-1K, consists of a single-protocol RNA-seq dataset comprising 1,035 samples. This dataset, named the **P**recision **R**NA-seq **E**xpression **C**ompendium for **I**ndependent **S**ignal **E**xtraction, encompasses 38% of publicly available high-quality RNA-seq data for *E. coli* K-12 and covers a wide range of growth conditions. The data were generated between 2013 and 2021.

To create the regulatory component of the resource, we employ ICA to extract 201 *i*ndependently *mod*ulated groups of genes called iModulons, which collectively capture 86% of known regulatory interactions. We showcase the utility of this resource by: (1) describing genome-wide expression patterns; (2) elucidating systems-level transcriptome properties and responses; (3) proposing new regulons for two putative transcription factors; (4) identifying the promoter sequence basis for two distinct subsets of the CRP regulon; (5) integrating an additional 1,675 high-quality publicly available *E. coli* K-12 samples and extracting similar regulatory modules; and (6) providing a workflow for system-level transcriptome analysis of external data using our

knowledge base.

The example workflow and all analyses presented in this study can be accessed and utilized through our GitHub repositories: `https://github.com/SBRG/precise1k-analyze` and `https://github.com/SBRG/precise1k`. The PRECISE-1K dataset, along with iModulons for Public K-12 and other organisms mentioned, can also be explored at `iModulonDB.org` [109].

PRECISE-1K serves as the expression component and iModulons serve as the regulation component of a comprehensive transcriptomic knowledge base. This resource empowers analyses that shed light on the transcriptomic responses of this critical model organism, enabling research in cellular biology, pathogenicity, and systems biology. Moreover, it offers valuable insights to inform the design of novel experimental studies. Beyond its applicability to *E. coli*, this resource provides a framework for extracting regulatory information in other organisms, particularly those lacking extensive prior annotation.

4.2 Results

4.2.1 PRECISE-1K is a 1,035-sample, high-precision, single-protocol RNA-seq compendium

To enable a comprehensive analysis of transcription and regulation in *E. coli* K-12 MG1655, we developed PRECISE-1K (Fig. 4.1A; Fig. C.1). PRECISE-1K is a large, curated expression compendium, comprising 1,035 RNA-seq samples generated by a single research group. The dataset follows a standardized experimental and data processing protocol (see Methods) and includes samples from 45 distinct projects. It encompasses a wide range of growth conditions, including: 5 strains, 4 temperatures, 5 pH levels, 9 base media, 18 carbon sources, 38 supple-

ments, 76 unique gene knockouts, 421 evolved samples, and 87 fed-batch cultures (Fig. C.2). The projects comprising PRECISE-1K involve adaptation to new growth conditions [110–114], expression of heterologous [115] and orthologous [116] genes, as well as a genome-reduced strain [46]. This compendium represents a significant expansion to the original 278-sample PRECISE [45], nearly quadrupling its size (Fig. 4.1B). Biological replicates exhibit strong correlation, with a median Pearson's r value of 0.99 (Fig. 4.1C). Therefore, PRECISE-1K provides a diverse set of conditions that allows for analysis of the *E. coli* transcriptome and its myriad responses.

Principal component analysis (PCA) of PRECISE-1K identifies some expected batch effects. The separation between samples in the principal component space is primarily driven by differences in growth conditions across projects (Fig. C.3). Projects involving diverse growth media (e.g., the two-component system knockout [117] and antibiotic resistance project [118]) and projects with significant genome modifications (e.g., genome-reduced E. coli strain [46]) exhibit notable divergence from other projects. Clustering by library preparer can largely be explained by project-based clustering, indicating that this commonly observed batch effect [100] is not prominent in PRECISE-1K; this observation is further supported by the strong correlations observed among biological replicates.

4.2.2 PRECISE-1K segments genes by expression, variance, and regulatory effect

We conducted a systems-level analysis of expression trends in PRECISE-1K to compare data-driven observations with prior expectations. Initially, we compared the median expression levels of genes across PRECISE-1K with their median absolute deviations (MAD) to establish expression-based categories for all genes (Fig. 4.1D). For instance, the glutamate-dependent acid

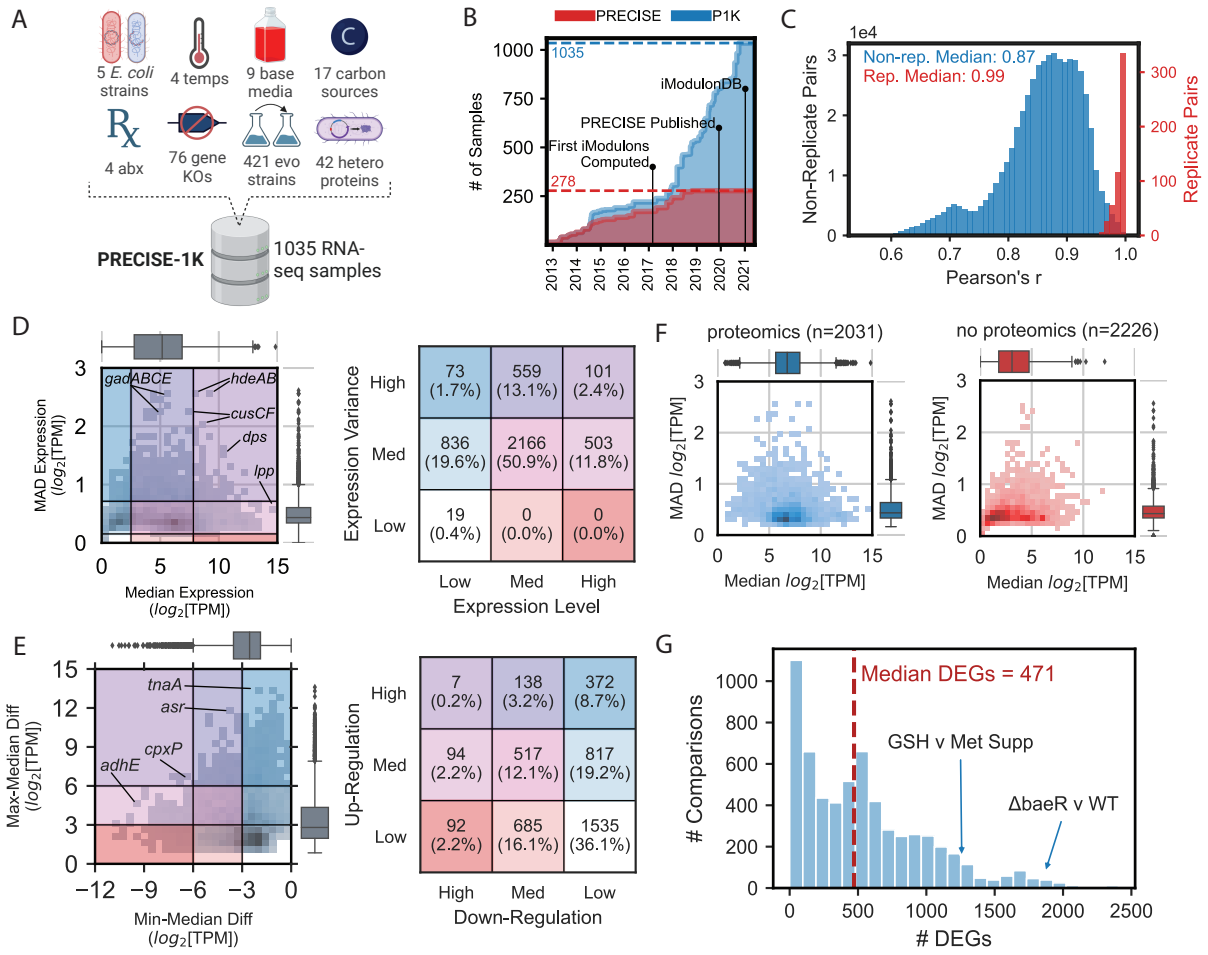


Figure 4.1: PRECISE-1K, a 1035-sample high-precision expression compendium, reveals expression trends in the *E. coli* transcriptome. **A)** Overview of construction of PRECISE-1K compendium. Values indicate the number of unique categories for each condition (except evo strains). abx = antibiotics. **B)** The growth in single-protocol transcriptomics samples contained in the PRECISE to PRECISE-1K databases. **C)** Histogram of Pearson's r for both all replicate pairs and all non-replicate pairs (pairwise combinations of samples across projects that are not direct biological replicates). Samples included in PRECISE-1K are required to have replicate correlations of at least 0.95. **D)** 2-D histogram of median expression level against median absolute deviation (MAD) of expression for all 4257 genes in PRECISE-1K. Table defines expression categories as per corresponding box color/location in histogram. For each axis, category splits are defined at median +/− 1 standard deviation. **E)** 2-D histogram of median-to-min expression difference against median-to-max expression difference for all 4257 genes in PRECISE-1K. Table defines regulatory categories as per corresponding box color/location in histogram. For each axis, low-to-medium split defined at 3 log_2[TPM] units (8-fold change from median expression); medium-to-high split defined at 6 log_2[TPM] units (32-fold change). **F)** Median vs MAD expression 2-D histogram, separated by availability of proteomics data in two large recent datasets [119, 120]. Blue = proteomics data available; red = no proteomics data available. **G)** Histogram of the number of differentially expressed genes (DEGs) computed between condition pairs within the same project (n=6103 pairs). GSH = glutathione, Met = methionine.

resistance system 2 genes (*gadABCE*) exhibit medium aggregate expression but display high variation across conditions due to the specifiticy of their response. On the other hand, the lipoprotein-encoding gene *lpp*, known for its abundance in *E. coli* [121, 122], exhibits the highest median expression with medium variation, likely owing to its structural role in peptidoglycan. The majority of genes demonstrate medium expression with medium variation, while only a small fraction of genes (101) show both high expression and high variability, including the copper/silver export system component *cusF*. Notably, 82% of genes (3505/4257) exhibit variation within one standard deviation of the overall median variation across all genes, while only 19 genes with low overall expression display low variation, mainly consisting of insertion elements and prophage genes.

Next, we compared the median expression levels of genes with their minimum and maximum levels to determine the extent of regulatory influence on expression level (Fig. 4.1E). Approximately 36.1% of genes exhibit a tight range of expression, indicating relatively low effects of regulation. However, 45.6% of genes display medium or high upwards inducibility, and 36% exhibit medium or high downwards inducibility, suggesting that regulatory effects can significantly impact gene expression levels for a majority of genes. For example, cpxP - a protein responding to extracytoplasmic stresses as part of the CpxAR two-component system [123] - has a nearly unique tendency to be both highly up- and down-regulated from its median level. This characteristic may result from CpxP's role as both a direct effector of various stress responses and a negative feedback regulator for the response pathway as a whole [124].

Additionally, PRECISE-1K sheds light on the relationship between gene expression and other data types. Genes with available proteomics data in two large datasets [119, 125] exhibit significantly higher expression, consistent with a known bias towards higher-expressed genes in

proteomics samples (Fig. 4.1F). However, no significant difference in variability was observed. Furthermore, poorly-annotated genes (referred to as the "y-ome" in E. coli [49]) have significantly lower expression compared to genes with more complete annotation, suggesting that the lack of transcription in standard laboratory conditions might contribute to the relative lack of annotation for these genes (Fig. C.4). Functional categories such as "Translation" and "Cell Cycle" exhibit the highest expression levels, while specialized categories like "Carbohydrate Metabolism" display lower median expression levels (Fig. C.5).

We conducted differential gene expression analysis within each project included in the PRECISE-1K compendium. Across all pairwise within-project comparisons, a median of 471 differentially expressed genes (DEGs) were identified (Fig. 4.1G). Some comparisons yielded minimal DEGs, while others resulted in a much larger number. For example, comparing wild-type growth in minimal media to the deletion of two-component system (TCS) response regulator *baeR* with ethanol supplementation yielded 1868 DEGs. Generally, gaining biological insights solely from DEGs may require analyzing hundreds to thousands of genes.

In summary, these findings demonstrate the ability of PRECISE-1K to capture genome-wide expression patterns, affirm existing expectations, and uncover new knowledge. Serving as an expression knowledge base, PRECISE-1K not only houses expression data but also enables knowledge-generating analyses. Quantifying the impact of regulation on gene expression at the systems level represents the next phase of knowledge extraction facilitated by this resource.

4.2.3 Top-down extraction of independently-modulated groups of genes captures the transcriptome at the systems level

We utilized Independent Component Analysis (ICA) to identify 201 iModulons from the PRECISE-1K dataset. iModulons are distinct groups of genes that co-vary across the dataset, captured by iModulon activity levels that represent their response in each PRECISE-1K condition. iModulons account for 83% of the total dataset variance. 117 iModulons are classified as Regulatory, exhibiting significant enrichment in genes belonging to known regulons (see 4.2A and Materials and Methods for regulatory enrichment details). These regulatory iModulons explain 56% of the total variance in PRECISE-1K. iModulons capturing smaller regulons closely align with known regulons, while those capturing larger regulons recover smaller subsets of the genes, leading to lower precision and recall (Fig. 4.2B).

Furthermore, 36 genomic iModulons capturing known genetic alterations and 17 biological iModulons composed of genes with shared functions but lacking significant regulon enrichment account for an additional 19% of the variance. 22 technical iModulons, explaining just 2% of the variance, are primarily dominated by a single short, uncharacterized gene, with 12 of them consisting of only that one gene. It is likely that these iModulons capture noise in the dataset. Moreover, nine uncharacterized iModulons collectively account for 6% of the variance in the dataset. Overall, 88% of the variance captured by iModulons can be attributed to either regulatory, genomic, or biological phenomena.

58% of genes (2485 out of 4257) are members of at least one iModulon. These genes exhibit higher expression variation compared to genes not present in any iModulons ($P = 1.03E-217$, Mann-Whitney U test, $m = 2485$, $n = 1772$) (see 4.2C). However, the median expression does not significantly differ ($P = 0.33$), indicating that iModulon membership is not limited to higher-

expressed genes. Interestingly, even 56% (823 out of 1473) of the less-expressed y-ome genes [49] are members of at least one iModulon, suggesting the potential of iModulons in uncovering putative functions for these uncharacterized genes.

The median Modulon consists of 10 genes, but many iModulons are much larger, such as global stress responses RpoS (122 genes) and SoxS (117) (Fig. C.6A). Among the 189 multi-gene iModulons, 77% (145) exhibit significantly intercorrelated genes, particularly among regulatory and biological iModulons (88% and 82%, respectively) (Fig. C.7). In contrast, genomic and technical iModulons show lower proportions of significantly intercorrelated genes (47% and 13%, respectively). Genomic iModulons, which capture genetic alterations in small subsets of the dataset, may not show global correlation, indicating that iModulons can capture localized expression patterns beyond global correlations. Interestingly, eight out of nine uncharacterized iModulons contain significantly intercorrelated genes, presenting opportunities for further biologically-relevant discoveries.

A substantial portion of genes in an iModulon (35%, 879 out of 2485) are members of two or more iModulons, with two genes (*ynfM* and *bhsA*) appearing in seven each (Fig. C.6B). Only 15% (131) of multi-iModulon genes are members of significantly correlated iModulons. However, within each of their iModulons, multi-iModulon genes rank in the 44th percentile in terms of intercorrelation with other iModulon genes. This suggests that multi-iModulon genes are influenced by distinct, recoverable signals, showcasing iModulons' ability to capture overlapping regulatory modules of varying scales. The relationships between iModulons and genes are concentrated in a subset of large iModulons and genes present in multiple iModulons (Fig. C.6C-D).

Moreover, 80 metabolism-related and 50 stress response-related iModulons account for

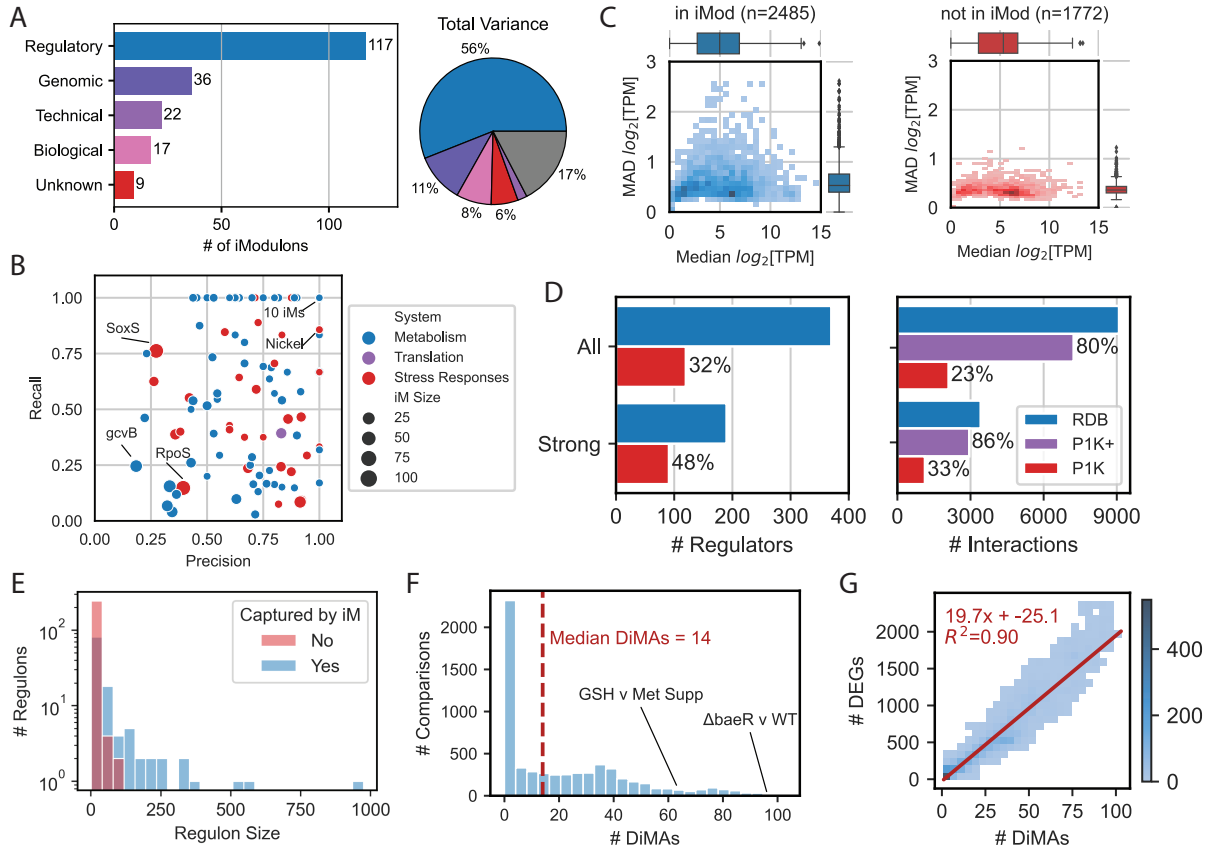


Figure 4.2: iModulons extracted from PRECISE-1K capture the transcriptional regulatory network. **A)** A breakdown of PRECISE-1K iModulons by their annotation category; see Methods for category details. Pie chart denotes iModulon annotation categories by percentage of variance explained. Gray wedge indicates variance unexplained by iModulons. **B)** Summary of precision and recall for 117 regulatory iModulons. RegulonDB (`http://regulondb.ccg.unam.mx`) [42] regulons used as reference. **C)** 2-D histograms of median gene expression and median absolute deviation in gene expression by iModulon membership. **D)** Comparison of regulators and regulatory interactions recovered by PRECISE-1K iModulons and available in RegulonDB. All = all evidence levels; Strong = only strong evidence interactions per RegulonDB; P1K+ = all interactions for which the corresponding regulator is captured by an iModulon. **E)** Histogram of RegulonDB regulon sizes, colored depending on whether each RegulonDB regulon is or is not captured by at least one PRECISE-1K iModulon. **F)** Histogram of the number of differential iModulon activities (DiMAs) computed between condition pairs within the same project ($n = 6103$; same as 4.1G). **G)** Comparison of number of DEGs and DiMAs for the same condition pairs. Linear best fit curve is shown in red, and indicates a 20-fold dimensionality reduction from DEGs to DiMAs. $n = 4483$ comparisons with non-zero DiMAs.

32% and 30% of the variance in PRECISE-1K, respectively (Fig. C.8A). This division emphasizes a "fear-greed" tradeoff, wherein metabolic capabilities are diversely regulated, while stress

responses are more centrally controlled. Notably, just two iModulons - RpoS and ppGpp, major stress response regulators - collectively account for 6% of the variance in the dataset (Fig. C.8B-C).

iModulons capturing signals of global regulators - defined here as those with more than 25 regulatory targets - contribute significantly to the overall dataset variance. For instance, flagella-related regulators FlhDC and FliA combinedly explain over 5% of the expression variance, while anaerobic growth regulators FNR and ArcA explain over 3% of the variance (Fig. C.8C). These findings underscore the ability of global regulators to mobilize large-scale transcriptomic responses, and they are responsible for the variance between wild-type control samples run across projects, despite overall tight correlation (Fig. C.9). Importantly, these batch variations are explicitly captured by iModulon activities.

4.2.4 Regulatory modules represent the majority of the known transcriptional regulatory network

iModulons extracted from PRECISE-1K reconstruct a substantial portion of the total regulatory interactions available in RegulonDB [42], the premier database for curated and validated regulatory network information for *E. coli*. Regulatory iModulons capture 32% of all known regulatory molecules (and 48% with strong evidence) (Fig. 4.2D). Furthermore, they reconstitute 23% of all specific regulatory interactions (33% of strong-evidence interactions). iModulons are known to capture regulatory signals by identifying the most strongly-regulated genes in a regulon based on promoter sequence [126]. This sequence-based effect likely accounts for the relatively lower precision and recall enrichment statistics observed for larger iModulons that capture more global regulators. Thus, considering a regulatory iModulon as a biomarker for all of its regula-

tor's interactions reveals that iModulons reconstitute 80% of all known regulatory interactions (86% when considering only strong evidence). Importantly, iModulons preferentially capture the signals of larger regulons (Fig. 4.2E), increasing their utility in describing the transcriptome state across growth conditions.

Subsampling PRECISE-1K and recomputing iModulons demonstrates regulatory network coverage at different compendium sizes. On average across five trials, 20%-scale subsamples of PRECISE-1K (207 samples) yield 111 iModulons, of which 67% (75) are regulatory iModulons also captured from PRECISE-1K (Fig. C.10A). As more samples are added, the total number of extracted iModulons increases; however, the relative fraction of regulatory iModulons decreases. Nonetheless, regulatory recovery increases with scale: 33% of strong-evidence regulators are captured in iModulons from 20%-scale subsamples, compared with 48% from PRECISE-1K's iModulons (Fig. C.10B). Captured regulatory interactions follow a similar pattern. Critically, the step from 80%-scale subsamples (828 samples) to full PRECISE-1K elicits an increase in regulatory discovery following a plateau between the 60%- and 80%-scales, indicating that PRECISE-1K's scale provides an advantage for regulatory recovery.

In all, iModulons provide the regulatory component of this transcriptome knowledge base. The subsequent sections demonstrate transcriptomic knowledge that can be derived from these regulatory modules.

4.2.5 Systems-level analysis of transcriptome states using regulatory modules

As iModulons explicitly represent activity levels, they facilitate the use of differential iModulon activity (DiMA) analysis. This type of analysis allows for a systems-level comparison of transcriptome states by reducing hundreds or thousands of differentially expressed genes (DEGs)

to a median of only 28 iModulons (Fig. 4.2F). When comparing any two conditions in PRECISE-1K, DiMA analysis yields nearly 20 times fewer differentially activated iModulons than DEGs (Fig. 4.2G), highlighting the particular utility of DiMA for systems-level transcriptional analysis. On average, DiMAs directly account for 37% of variance between conditions. Considering that all iModulons together explain a median of 80% of variance between conditions, DiMAs contribute to a median of 47% of the variance explained by all iModulons (Fig. C.11).

The activities of iModulons reflect the overall activity state of a transcriptional regulator across environmental conditions in PRECISE-1K. A stimulon is a higher-level regulatory structure composed of multiple regulons that respond to a particular stimulus (Fig. C.1). While iModulons encompass independently modulated groups of genes, in many cases, these independent groups of genes are regulated in response to similar environmental stimuli, forming a stimulon. Two-component systems (TCS), consisting of a membrane-bound sensor and a cytoplasmic response regulator, enable the cell to sense and respond to important extracellular signals. iModulons derived from PRECISE-1K capture the response signals for 15 of the 27 known TCS response regulators, providing insight into the cell's regulatory response to critical stimuli such as nitrogen, inorganic phosphate, and alkali metals.

Furthermore, iModulons can be clustered based on their activities, revealing higher-order structures in the *E. coli* transcriptome. For instance, one cluster captures the joint regulation of flagella formation by the transcription factor complex FlhDC and the sigma factor FliA (σ_{28}) (Fig. C.12). Six iron-related iModulons, five anaerobiosis-related iModulons, and four amino acid-related iModulons also group together based on their activities. Thus, the combination of iModulons can shed light on broad transcriptome patterns, providing a new definition of a stimulon.

4.2.6 Regulon discovery for putative transcription factors YgeV and YmfT

Functional annotation for putative transcription factors (TFs) remains a challenging task [127–129]. However, iModulons present a powerful tool for discovering and analyzing new regulons. PRECISE successfully elucidated the regulons for three previously uncharacterized TFs (YieP, YiaJ/PlaR, and YdhB/AdnB) and expanded the regulons of three known TFs (MetJ, CysB, and KdgR) [45]. Many of these regulatory interactions were further confirmed through DNA-binding profiles [45, 130, 131]. Additionally, iModulons derived from a microarray dataset predicted three novel regulons [132]. The iModulons from PRECISE-1K reaffirm these previous findings and reveal two new potential regulons.

The putative YgeV regulon includes 13 genes, among which 7 are implicated in nucleotide transport and metabolism (Fig. 4.3A). YgeV is predicted to be a Sigma54-dependent transcriptional regulator, and Sigma54-dependent promoters were previously identified upstream of the xdhABC and ygeWXY operons, which are part of the YgeV iModulon [133]. Although the iModulon does not contain the gene ygeV, ygeV is transcribed divergently from ygeWXY. A prior study [134] indicated reduced expression of ygfT in a YgeV mutant strain. As ygfT is part of the YgeV iModulon, this suggests that YgeV may serve as an activator for the genes within its iModulon. The activity of the YgeV iModulon rarely deviates from the reference condition; however, it is most active when TCS response regulators BaeR or CpxR are knocked out and the strain is exposed to ethanol (Fig. 4.3B). Therefore, we hypothesize that the TF YgeV responds (either directly or indirectly) to ethanol to activate genes related to purine catabolism and is repressed by TCS BaeRS and CpxAR.

The putative YmfT regulon consists of 15 genes, including *ymfT* itself. It includes 12 out of the 23 genes in the e14 prophage [135] (Fig. 4.3C). The putative YmfT iModulon shows

the highest activity in strains lacking the ferric uptake regulator Fur or when challenged with oxidative stress induced by hydrogen peroxide (Fig. 4.3D). The absence of Fur leads to overproduction of iron uptake proteins, oxidative damage, and subsequent mutagenesis [136]. Hence, we propose that YmfT responds to oxidative stress to modulate the expression of the e14 prophage.

These examples demonstrate the potential of iModulons in predicting new regulons and identifying optimal conditions for studying their activities.

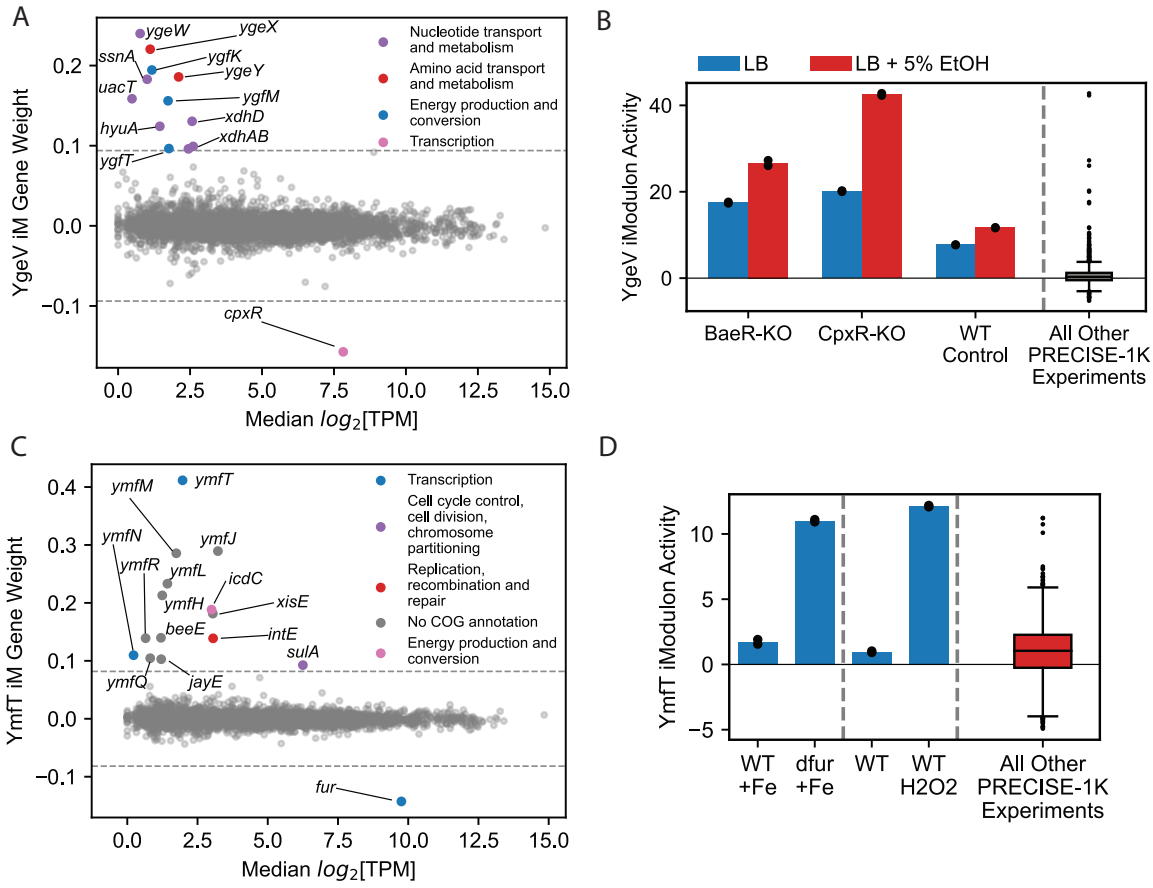


Figure 4.3: iModulons discover new regulons. **A)** iModulon gene weights for the putative YgeV iModulon vs. median log2[TPM]. **B)** Activity of the YgeV iModulon in different media conditions. Each colored bar is the mean of two biological replicates (shown as individual black points). **C)** iModulon gene weights for the putative YmfT iModulon vs. median log_2[TPM]. **D)** Activity of the YmfT iModulon in different media conditions. Each colored bar is the mean of two biological replicates (shown as individual black points).

4.2.7 Stratifying promoter-level mechanisms of Crp regulation

iModulons uncover distinct, independent sub-groups of genes within global regulons, revealing unique regulatory dynamics. An illustrative example is seen in the Fur-1 and Fur-2 iModulon activities, which each capture subsets of the Fur regulon. These activities exhibit non-linear correlations based on both iron availability and aerobicity [118].

In this section, we investigate how iModulons reflect the biochemical mechanisms of transcription factor (TF) binding by examining the relationship between two iModulons - Crp-1 and Crp-2 - that stratify the CRP regulon. The CRP regulon contains multiple RNA polymerase-interacting domains (Ar1-3) [137] that facilitate its binding to Class I and Class II promoters. Class I promoters involve binding centered 61.5 base pairs upstream of the transcription start site, while Class II promoters are centered 41.5 base pairs upstream [138, 139] (Fig. 4.4A).

The activities of the Crp-1 and Crp-2 iModulons across all PRECISE-1K conditions exhibit a distinct nonlinear relationship (Fig. 4.4B). As expected, low activities of both iModulons correspond to the deletion of CRP, which is known to activate most of the genes in these two iModulons. Notably, the deletion of the Ar2 binding domain - implicated in Class II regulation - results in some Crp-1 activity but no Crp-2 activity (orange dot in 4.4B). Additionally, CRP binding sites for genes unique to Crp-1 are broadly distributed, while Crp-2-specific genes have CRP binding sites more consistently at the Class II location (Fig. 4.4C). A steady-state biophysical model, incorporating 10-fold different binding affinities for Class I and Class II binding sites, produces a similar binding site occupancy relationship as observed between the Crp iModulon activities (Fig. 4.4D). Based on this evidence, we propose that the Crp-1 and Crp-2 iModulons correspond to Crp regulatory activity at Class I and Class II promoter genes, respectively. This analysis underscores the capability of PRECISE-1K iModulons to capture multi-dimensional

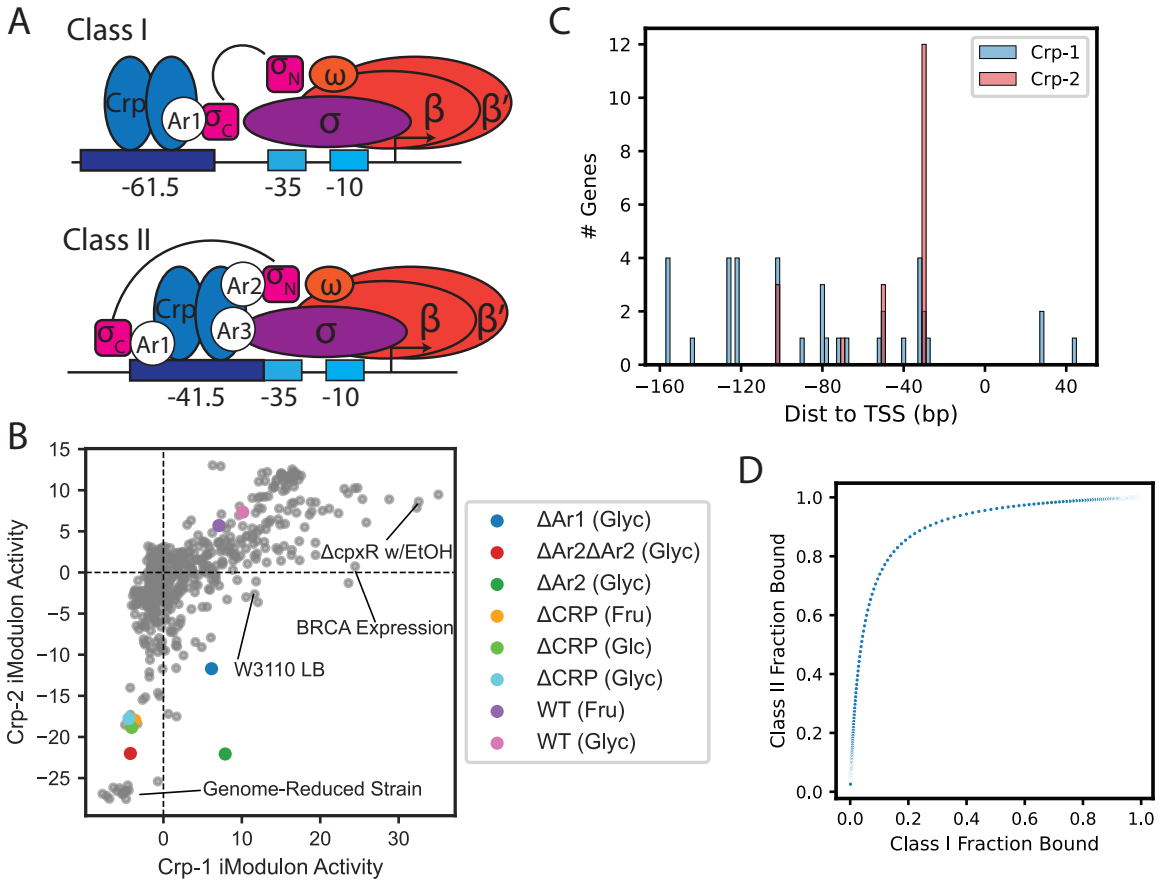


Figure 4.4: iModulons stratify existing regulons by mode of binding. **A)** Diagram of Class I and Class II CRP promoters. Arrow indicates transcription start site. σ = RNA polymerase (RNAP) sigma factor; σ_{NC} = sigma factor N- and C-terminal regions; Ar1-3 = CRP activating regions (RNAP interaction sites). **B)** iModulon phase plane between Crp-1 and Crp-2 iModulons. Colored points from samples involving partial and total CRP deletions. Ar regions correspond to panel A. Glyc = glycerol carbon source; fru = fructose; glc = glucose. **C)** Histogram of CRP binding site locations for Crp-1 and Crp-2 iModulons. TSS = transcription start site of transcription unit for each gene. Data from RegulonDB. **D)** Simulated binding curve for CRP Class I and Class II promoters. Each point indicates a particular CRP concentration. Binding modeled as 10x tighter at Class II vs Class I promoters.

regulatory effects within a single regulon.

4.2.8 Incorporating 1,675 high-quality publicly-available transcriptomes into the knowledgebase highlights method's scalability and robustness

To further expand our dataset, we sourced all publicly-available RNA-seq data for *E. coli* strain K-12 from NCBI's Sequence Read Archive (SRA) [140]. From 3,230 K-12 samples, our processing and quality control pipeline yielded 1,675 high-quality K-12 expression profiles. We combined these samples with PRECISE-1K to yield the "K-12 Dataset," a high-quality transcriptomics dataset consisting of 2,710 expression profiles (Fig. 4.5A). These profiles come from 134 different projects, including 15 K-12 substrains and 9 distinct temperatures and pHs. ICA decomposition of the K-12 Dataset yields 194 iModulons.

The distribution of iModulons by category – both in number and by explained variance – is broadly similar to that of PRECISE-1K. Regulatory iModulons account for 64% of the total number, and 57% of the total variance in the dataset (Fig. 4.5B). Coverage of known regulatory network interactions increases only minutely as compared with PRECISE-1K alone, despite the more than doubling of the dataset's size (Fig. 4.5C). Indeed, 89% of K-12's explained variance comes from 155 iModulons highly correlated with iModulons extracted from either PRECISE or PRECISE-1K (Fig. 4.5D). In contrast, 45% of explained variance from PRECISE-1K comes from 134 iModulons not present in PRECISE. Nonetheless, 67 iModulons captured in the original PRECISE are retained in both PRECISE-1K and K-12, accounting for sizable fractions of explained variance in each of the latter datasets. The iModulon structure remains largely consistent as dataset scale is increased; in general, higher-variance signals discovered by smaller-scale datasets are supplemented with new, more niche iModulons, rather than the entire iModulon structure shifting with scale. iModulons can also explain a slightly larger fraction of variance in PRECISE-1K than in the K-12 Dataset. iModulons extracted from just the 1,675 publicly-

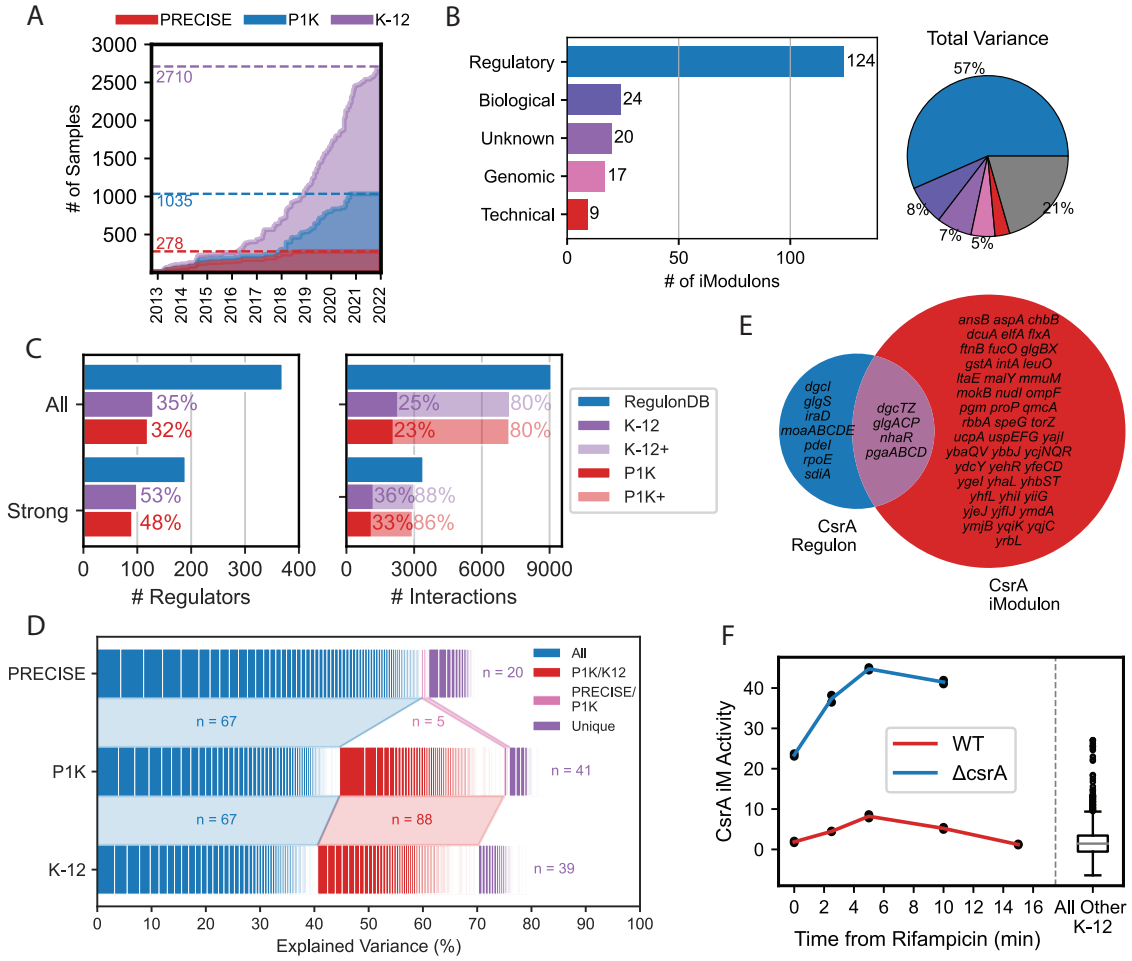


Figure 4.5: Adding public K-12 data to PRECISE-1K highlights P1K's stability. **A)** K-12 is a combined dataset with P1K (1035 samples) plus all publicly-available high-quality RNA-seq data for *E. coli* K-12 (1675 samples). A) Growth of high-quality RNA-seq data for K-12. **B)** K-12 iMs by their annotation category (see 4.2A legend). **C)** Comparison of regulators and regulatory interactions recovered by K-12 and available in RegulonDB. All = all evidence levels; Strong = strong evidence per RegulonDB; K-12+ = interactions for which regulator is captured by K-12 i. P1K values from 4.2D included for comparison. **D)** Comparison of iMs from three RNA-seq datasets: PRECISE [45]; P1K (this paper); and public K-12. Each small rectangle is an iM. Pairwise Pearson correlations were performed between PRECISE and P1K iMs, and between P1K and K-12 iMs; iMs with correlations over 0.3 were considered to be the same iM (median PRECISE/P1K r is 0.68; P1K/K-12 0.70). Blue = all 3 datasets; pink = only PRECISE/P1K; red = P1K/K-12 only; purple = unique to dataset. Explained variance within each dataset. iMs ordered by which dataset(s) they appear in, and sorted in decreasing order of explained variance within each dataset appearance category. **E)** Overlap between the CsrA regulon per RegulonDB and the CsrA iM. **F)** Activity of the CsrA iM after arrest of transcription initiation via addition of rifampicin (data from Potts et al [141]).

available K-12 samples are similar to those extracted from the 2,710-sample compendium, albeit with lower regulatory recovery (Fig. C.13). Taken together, these results suggest that PRECISE-1K has sufficient scale and condition variety to represent the *E. coli* TRN, and additions of data beyond this scale may provide diminishing returns.

However, specific conditions in the K-12 dataset enable regulatory discovery. For example, 18 samples from a project exploring the post-transcriptional carbon storage regulator CsrA regulon [141] enabled recovery of a CsrA iModulon that is unique to the K-12 dataset. The CsrA iModulon is much larger than the known CsrA regulon: it contains 65 genes, of which 10 overlap with the 21-gene CsrA regulon (Fig. 4.5E). Nonetheless, the enrichment of CsrA regulon genes in the iModulon is significant (adjusted $P = 6.7E-9$), and the genes in both the iModulon and regulon are particularly highly weighted in the iModulon. Moreover, the iModulon is much more highly active in a CsrA deletion strain after arrest of transcription initiation than the wild-type strain or other K-12 samples (Fig. 4.5F), indicating relief of CsrA repression. Thus, the genes unique to the iModulon are candidates for expansion of the CsrA regulon.

4.2.9 Applying the knowledge base to new data: the anaerobic to aerobic transition

This knowledge base enables the analysis of new E. coli RNA-seq datasets. To demonstrate this capability, we utilized one project from the public K-12 Dataset called AAT (anaerobic-aerobic transition). The AAT project captured six time-points in triplicate, ranging from 0 to 10 minutes after aeration of a previously anaerobic chemostat culture of E. coli K-12 W3110 [142]. By inferring PRECISE-1K iModulon activities for the AAT project, there was no need for full re-processing through the entire workflow. This inferred information allowed us to analyze AAT's

samples both within the project and in the context of all PRECISE-1K's samples. The code used for this case study is available at `https://github.com/SBRG/precise1k-analyze` and can be utilized for the analysis of any new data.

Our hypothesis was that certain iModulons would respond to the onset of aerobic growth (Fig. 4.6A). For instance, the regulators Fnr and ArcA are both influenced by oxygen availability. Fnr is activated upon acquiring an iron-sulfur (4Fe-4S) cluster and dimerizing, while oxygen inactivates Fnr by oxidizing the iron-sulfur cluster [143–146]. Fnr's active state leads to the activation of anaerobic metabolism genes and repression of aerobic metabolism genes [147]. On the other hand, ArcA is the transcription factor component of a quinone-sensing two-component system. During aerobic growth, quinols are oxidized to quinones as part of the electron transport chain, which prevents the sensor kinase ArcB from phosphorylating and activating ArcA [148, 149]. ArcA predominantly represses aerobic metabolism genes while also activating a few fermentative genes [150–152]. Additionally, several aerobic metabolism genes, especially those encoding oxidoreductases and electron transport chain components, require iron-sulfur clusters to function. Consequently, the global iron regulator Fur, which represses iron acquisition genes when bound to iron, is also involved in this transition [153, 154]. Lastly, oxidative phosphorylation under aerobic conditions generates reactive oxygen species (ROS), which triggers the SoxS and OxyR responses [155, 156].

Identifying the iModulons with divergent activities in AAT compared to the rest of PRECISE-1K revealed several iModulons related to energy metabolism. For example, the formate hydrogen lyase (FHL) iModulon displayed a maximum absolute activity in AAT that was six standard deviations away from the PRECISE-1K median. FHL is known to be active under anaerobic conditions during glucose fermentation [157].

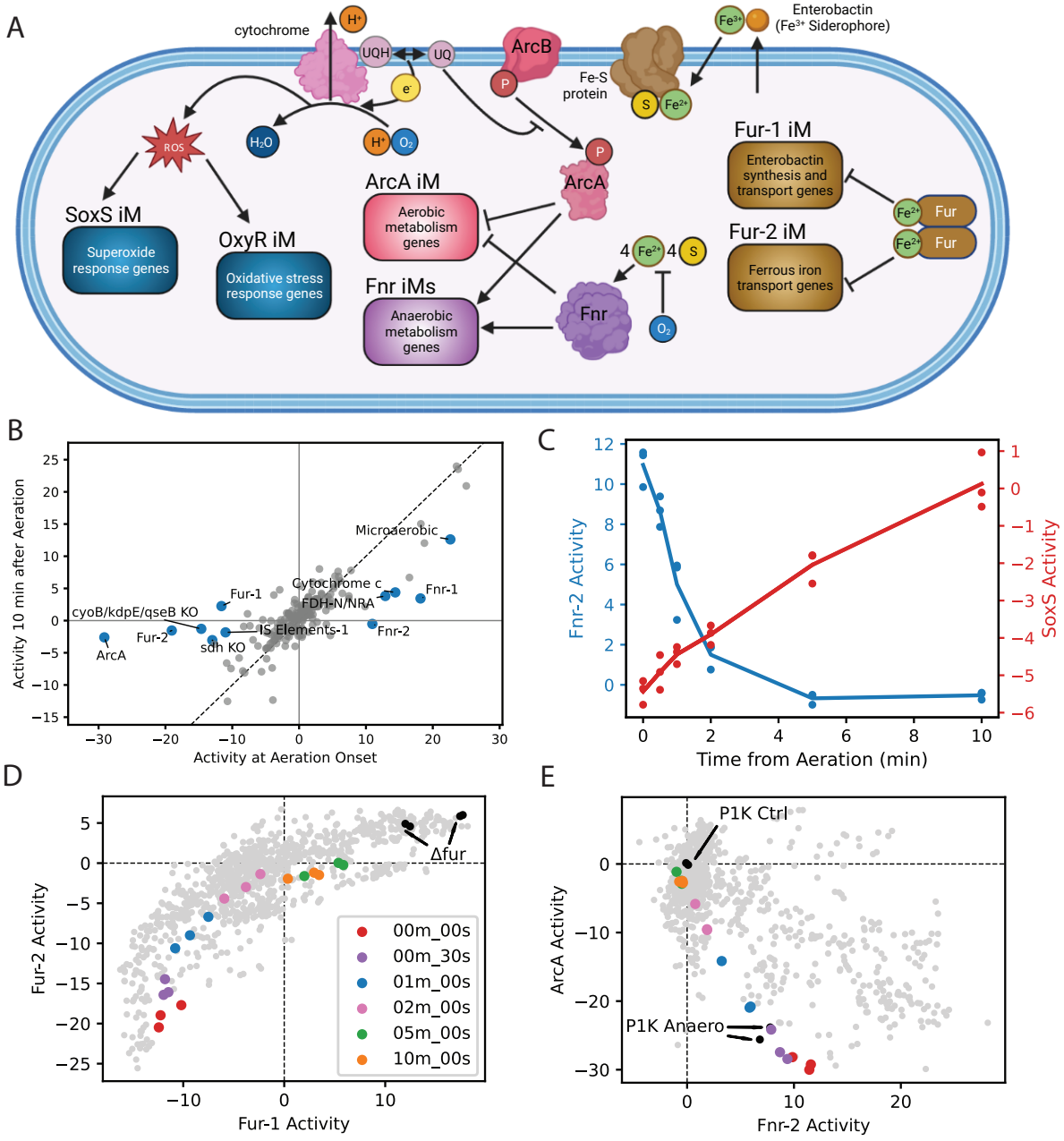


Figure 4.6: PRECISE-1K and iModulons for new data analysis. Example new data from AAT [142] (anaerobic-aerobic transition) (from public K-12 data): 6 timepoints from 0 to 10 minutes after aeration of anaerobic chemostat. **A)** Selected iMs and systems involved in aerobic transition. **B)** Differential iM activity (DiMA) plot comparing iM activities at aeration onset and 10 minutes after aeration. iMs with significant activity differences between the two time points are in blue and labeled (see Methods for DiMA details). **C)** iM activity by time from aeration for Fnr-2 and SoxS iMs. **D)** Phase plane comparing activities of Fur iMs for all PRECISE-1K samples (gray) and aat samples (colored). Black dots = fur KO. **E)** Phase plane comparing activities of Fnr-2 and ArcA iMs. aat color scheme same as **D**.

For further characterization of iModulon activity changes within AAT, DiMA analysis helped identify iModulons that changed significantly between any two sets of samples. Comparing aeration onset to 10 minutes post-aeration highlighted the roles of key energy metabolism global regulators in facilitating this transition (Fig. 4.6B). Fnr was more active at onset, while ArcA and Fur exhibited significantly increased activity 10 minutes after aeration. Fnr's activity decreased nonlinearly following aeration of the culture, reaching its aerobic growth reference level within 5 minutes (Fig. 4.6C). Conversely, SoxS iModulon activity increased as aeration proceeded. Activity clustering revealed increased activity of the anaerobic stimulon at aeration onset, followed by increased activation of the iron stimulon 10 minutes post-aeration (Fig. C.14).

Activity phase planes, which compare two iModulons' activities across conditions, proved to be another valuable tool for analyzing new data. The dynamic transcriptomic changes in the AAT project were particularly evident in the Fur-1/Fur-2 (Fig. 4.6D) and Fnr/ArcA (Fig. 4.6E) phase planes. As aerobic metabolism took over, iron-related genes repressed by Fur during anaerobiosis increased in activity as the demand for iron increased. Simultaneously, the activity of the anaerobic regulator Fnr decreased as the aerobic regulator ArcA's activity increased, and both eventually reached activity levels similar to PRECISE-1K's aerobic growth control condition 10 minutes after aeration.

Overall, these observations shed light on the essential systems-level changes in the transcriptome composition during the anaerobic-aerobic transition and demonstrate the utility of PRECISE-1K as an analysis resource. Furthermore, they showcase the comprehensive interpretation of TRN functions accomplished through the use of iModulon activity phase planes.

4.3 Discussion

This study establishes a multi-scale gene expression and regulation knowledge base for *E. coli*. The expression component is PRECISE-1K, a single protocol, high quality RNA-seq dataset containing 1,035 samples covering a wide range of growth conditions. PRECISE-1K enables genome-wide categorization of genes based on expression level and expression variance across conditions. Using machine learning, we recover 117 regulatory modules (iModulons) from PRECISE-1K that reconstitute 86% of known regulatory interactions. iModulons - unlike principal components - explain variance in terms of TRN, not statistical magnitude. PRECISE-1K and its iModulons constitute the most complete top-down, computational transcription and regulation knowledge base yet generated for a microorganism. This resource enables regulon discovery and empowers novel experimental design. Most importantly, this resource empowers deep systems-level analysis of novel data.

We demonstrate that iModulons capture fundamental regulatory modes, not dataset-specific artifacts. iModulons from PRECISE-1K capture nearly all of the regulatory iModulons extracted from earlier version PRECISE. Increasing the dataset size nearly four-fold does not hinder regulatory discovery; in fact we more than double the number of discovered regulatory iModulons. Conversely, decreasing the dataset's scale via subsampling yielded poorer regulatory recovery. This potential highlights the central role that top-down, data-driven methods must take in transcriptional regulatory discovery across organisms. Indeed, iModulons have already successfully generated top-down regulatory information for other organisms [45, 104–108, 158]. Continued expansion of RNA-seq datasets for these and new organisms will likely drive further regulatory discovery.

iModulon activities enable analysis of the functional transcriptome under specific en-

vironmental or genetic conditions. We demonstrate this capability by capturing two different functional regulatory modes of the Crp regulon based on binding site location. DiMA analysis also greatly simplifies differential expression analysis; with an average of nearly twenty times fewer significantly differential variables to analyze, DiMA analysis empowers systems-level analysis of transcriptomic changes, as demonstrated in the anaerobic-aerobic transition case study.

Critically, PRECISE-1K and iModulon activities enable us to discover and partially characterize putative regulons for predicted transcription factors. We demonstrate this capability by assigning a putative function in ethanol stress tolerance related to nucleotide metabolism to the YgeV regulon, based on the YgeV iModulon activation pattern. In particular, this activation coincides with knockouts of two-component system response regulators BaeR and CpxR; thus, YgeV's role in nucleotide metabolism upon ethanol stress response may arise as a compensatory mechanism following inactivation of these more prominent TCS regulators. The specificity of this activating condition may play a role in explaining why the functions of this regulator and the genes in its regulon remain unknown. Indeed, iModulons have already proven useful in studies to characterize regulators and their regulons [130, 131, 159]. PRECISE-1K and the Public K-12 dataset likely contain other instances of untapped insights and should continue to be mined for such discoveries.

However, we also highlight the need for judicious selection of growth conditions to maximize potential for regulatory elucidation. When we added all high-quality public K-12 data to PRECISE-1K, the iModulon structure remained quite similar, with the K-12 Dataset's 124 regulatory iModulons accounting for 88% of known TRN interactions. This result highlights two key points. Firstly, PRECISE-1K has sufficient scale and diversity to enable broad TRN discovery while avoiding noise introduced by combining data from multiple sources. Secondly, adding large

numbers of RNA-seq samples beyond the scale of PRECISE-1K can yield diminishing returns. That said, certain specific new conditions from the K-12 Dataset were disproportionately useful - for example, a project perturbing the CsrA regulator enabled extraction of a corresponding regulatory iModulon which suggests expansion of the CsrA regulon. These observations likely highlight a limitation in the diversity of the available data, rather than of iModulons themselves. Thus, capturing additional unrecovered regulatory signals will likely rely on selection of growth conditions that activate niche transcriptional regulators with small regulons. Indeed, PRECISE-1K and the K-12 Dataset provide a blueprint for which conditions to prioritize for future discovery. Our knowledge base provides a centralized reference for assessment of gene expression and regulatory activity across conditions, empowering prudent study design. This capability is especially important for cost-, labor-, or time-intensive experiments, such as proteomics or metabolomics.

Our example analysis of the AAT project from the K-12 Dataset demonstrates perhaps the most exciting application of PRECISE-1K: analysis and contextualization of new RNA-seq datasets. PRECISE-1K's iModulons clearly capture and summarize the regulatory dynamics at play during aerobic metabolism transition. We provide a variety of tools, both here and in our previously published code package [160] that will easily facilitate similar analyses for any other dataset. In this way, PRECISE-1K is not just useful in and of itself but as a backdrop for deriving regulatory insight from new data. Our example workflow for analyzing new data with PRECISE-1K is available at at `https://github.com/SBRG/precise1k-analyze`; all other analyses from this paper are available for use at `https://github.com/SBRG/precise1k`. These analyses have already enriched multi-omic studies of the aerobic respiration system [161], the adaptation of different E. coli strains [162], and the response of E. coli to antibiotics [163].

Overall, PRECISE-1K and iModulons represent a critical resource for studying expression

and regulation in E. coli. We believe this resource should be a standard tool for systems-level analysis of E. coli RNA-seq data from all sources. As the number of publicly available datasets increases for other microorganisms, this study serves as a roadmap for interrogating similar datasets for less characterized organisms, with the potential to yield equally impactful insights into those organisms' transcription and regulation characteristics. PRECISE-1K is disseminated through `iModulonDB.org`.

Limitations of the study

Although this resource presents many opportunities, some limitations also merit mention. First of all, assembling at least 200 high-quality, single-protocol RNA-seq profiles presents an up-front challenge for generating a PRECISE database for other organisms. While combining publicly available data can help, we have demonstrated that single-protocol datasets provide more regulatory elucidation on a per sample basis. Secondly, while PRECISE-1K does contain a broad range of growth conditions, this set is by no means exhaustive. Thus, minimal expression and regulatory knowledge can be provided for these missing conditions. Thirdly, iModulons are subject to limitations due to the ICA algorithm by which they are computed. For example, ICA assumes that each iModulon results from a single signal (regulator); therefore, genes with multiple regulators - or complex, multi-regulator regulons - can be more difficult to capture in iModulons. Also, ICA does not allow for hierarchy; thus, iModulons do not always capture the effects of regulators on other regulators, i.e. the activation of a set of local regulators by a global regulator. Additionally, while ICA does maximize statistical independence, its components are still based somewhat on variance. Thus, as dataset scale increases, signals that were captured from smaller datasets may not be captured from larger datasets because they account for relatively

less variance in the larger dataset (hence the observation of some unique PRECISE iModulons). iModulons also cannot directly capture a true TRN - iModulons represent groupings of genes whose expression signals are intercorrelated while independent from other genes, not groupings of genes directly influenced by a regulator (as iModulons are computed without any prior knowledge of the TRN). Finally, DiMas - while indispensable for systems-level regulatory analysis - are not guaranteed to capture all individual gene-level changes in a given comparison. Indeed, the range of variance explained by DiMAs for any given condition comparison is wide; although a median of 47% of variance is explained by DiMAs, many comparisons are much more "lossy" than this. Thus, it remains important to analyze gene expression data directly - for which PRECISE-1K itself may be used. These important caveats should be kept in mind when using this resource to analyze new data or analyzing this resource itself.

4.4 Methods

RNA sequencing

3 mL of cell broth (OD600 of about 0.5, unless otherwise specified in sample metadata file) was immediately added to two volumes Qiagen RNA-protect Bacteria Reagent (6 mL), vortexed for 5 s, incubated at room temperature for 5 min, and immediately centrifuged for 10 min at 11,000×g. The supernatant was decanted, and the cell pellet was stored in the -80 °C. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 min at 37 °C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit (Cat. no. 74104) columns and following vendor procedures. An on-column DNase-treatment was performed for 30 min at room temperature. RNA was quantified using a

Nanodrop and quality assessed by running an RNA-nano chip on a bioanalyzer. The rRNA was removed using Illumina Ribo-Zero rRNA removal kit (Cat. no. 20037135) for Gram-negative bacteria. A KAPA stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used following the manufacturer's protocol to create sequencing libraries with an average insert length of around $\tilde{3}$00 bp. Libraries were run on a HiSeq4000 or NextSeq (Illumina).

RNA-seq processing and quality control

Starting from 1055 candidate samples, data was processed using a Nextflow [164] pipeline designed for processing microbial RNA-seq datasets [160], and run on Amazon Web Services (AWS) Batch.

First, raw read trimming was performed using Trim Galore (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) with the default options, followed by FastQC (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) on the trimmed reads. Next, reads were aligned to the E. coli K-12 MG1655 reference genome (RefSeq accession number NC_000913.3) using Bowtie [165] with the following non-default options: -X 1000, -3 3, -n 2. The read direction was inferred using RSEQC [166] before generating read counts using featureCounts [167] with the following non-default options: -p -B -C -P -fracOverlap 0.5. Finally, all quality control metrics were compiled using MultiQC [168] and the final expression dataset was reported in units of log_2-transformed Transcripts Per Million (log_2[TPM]).

Samples were considered "high-quality" if they met all of the following criteria: - "Pass" on the all of the following FastQC checks: `per_base_sequence_quality`, `per_sequence_quality_scores`, `per_base_n_content`, `adapter_content` - At least 500,000 reads mapped to coding sequences (CDS) from the reference genome (NC_000913.3) - Not an

outlier in hierarchical clustering based on pairwise Pearson correlation between all samples (outlier defined as cluster with number of samples ¡1% of the total number of samples) - Minimum Pearson correlation with biological replicates (if any) 0.95 (if more than 2 biological replicates, keep samples with high correlation in "greedy" manner, dropping samples that have at least one sub-threshold correlation with all other replicates)

Short non-coding transcripts (¡100 nucleotides) and extremely low-expression transcripts (FPKM ¡ 10) were also removed to reduce noise.

Following this processing and QC workflow, 1,035 high-quality RNA-seq samples (each with 4,257 gene expression measurements) remained. These samples and their metadata define PRECISE-1K. log_2[TPM], raw read count, QC data files, and sample metadata for all 1,055 original samples may be found in the data directory of this project's GitHub repository.

Differentially expressed gene (DEG) computation

Differentially expressed genes (DEGs) were identified using the DESeq2 package [169] on the PRECISE-1K RNA-seq dataset. Genes with a log_2 fold change greater than 1.5 and a false discovery rate (FDR) value less than 0.05 were considered to be differentially expressed genes. Genes with p-values assigned "NA" based on extreme count outlier detection were not considered as potential DEGs. The number of DEGs was computed for each unique pair of conditions within each project in PRECISE-1K, for a total of 6104 pairwise computations.

iModulon computation

log_2[TPM] data (4257 gene rows by 1035 sample columns) was centered to the control condition (log-phase growth in M9 minimal media with glucose; sample IDs `"p1k_00001"` and

"p1k_00002"); the mean log_2[TPM] of these two samples was computed, and the resultant 4257-gene log_2[TPM] vector was subtracted from all 1035 samples (columns) of the log_2[TPM] data table (including the control samples themselves, such that the mean of these samples was equal to 0).

No batch effect correction method (such as ComBat-Seq) was used - use of such methods significantly reduced regulatory signal discovery in testing. Many common types of batch variation - e.g. temperature, pH, growth phase - mediate expression changes through the TRN anyhow. Thus these minute perturbations - along with much larger variation across samples and projects - initiate the variant signals needed for ICA to identify regulatory activity.

The Scikit-learn implementation of FastICA [170] was used to run ICA on the centered log_2[TPM] data table. FastICA numerically solves the matrix decomposition equation $\mathbf{X} = \mathbf{MA}$; $\mathbf{X}$ is the input matrix; $\mathbf{M}$ is the "iModulon" matrix, and $\mathbf{A}$ is the "Activity" matrix; these terms will be used from here on in lieu of the traditional terminology $\mathbf{X} = \mathbf{SA}$ ("signal" and "mixing" matrices) to avoid confusion with the stoichiometric matrix $\mathbf{S}$ from metabolic modeling. In this context, $\mathbf{M}$ has dimensions of number of genes by number of components, and $\mathbf{A}$ has dimensions of number of components by number of samples. Thus, the $\mathbf{M}$ matrix contains weightings that specify how much each gene (row) belongs to each independent component (IC; column). The $\mathbf{A}$ matrix contains weightings that indicate how active each IC (row) is in each sample (column).

Unlike PCA, this method requires pre-specification of the number of components (parameter name `n_components`; also known as dimensionality) to use (the number of columns in $\mathbf{M}$ and number of rows in $\mathbf{A}$). In order to choose an optimal dimensionality, the previously described OptICA method [171] was used.

For PRECISE-1K, the selected optimal dimensionality by this method was 290. The

robust **M** and **A** matrices from this dimensionality run were selected, yielding 201 ICs. Thus, the final **M** matrix has dimensions of 4257 genes by 201 independent components, and the final activity matrix **A** has dimensions of 201 iModulons by 1035 samples.

The **M** matrix contains gene weightings, indicating how much each gene (row) "belongs" to each component (column), with larger absolute values indicating more association of a particular gene with a particular IC. For a given IC, gene weightings are mostly normally distributed around 0, with a few outlier gene weightings deviating from 0. To define an iModulon, a cutoff must be defined that allows segmentation of the genes in an IC based on their gene weightings. These cutoffs were determined with a previously described method [45] using D'Agostino's K2 test for normality. In this way, the final 201 iModulons were computed from the 201 independent components. A binary matrix $\mathbf{M_{binary}}$ was then constructed with the same dimensions as **M**; for a given gene (row)/iModulon (column) entry, a 1 indicates membership of the gene in the iModulon, and a 0 indicates that the gene is not a member of the iModulon.

The final matrices **M**, $\mathbf{M_{binary}}$ and **A** (along with iModulon membership thresholds as defined above, regulatory annotation as described below, and all other iModulon metadata) are available in the supplementary data files and in this project's GitHub repository.

iModulon annotation and curation

Using the gold-standard TRN reference annotation downloaded from RegulonDB v10.5 [42], enrichment of the set of genes in each iModulon against known RegulonDB regulons was computed using Fisher's Exact Test, with false discovery rate controlled at 10^{-5} using the Benjamini-Hochberg correction. By default, iModulons were compared to all possible single regulons and all possible combinations of two regulons (intersection only). The regulons used

by default consisted of only strong and confirmed evidence regulatory interactions, per RegulonDB. When multiple significant enrichments were available, the enrichment with the lowest adjusted P value was used for annotation. In the event of near equal P values (within an order of magnitude) across multiple enrichments, the priority was given to intersection regulons, followed by single regulons, followed by union regulons. If no significant enrichments were available, the following adjustments were used, in this order: relax evidence requirement to include weak evidence regulatory interactions; search only for single regulon enrichments; allow up to 3 regulons to be combined for enrichment; allow regulon unions as well as intersections (with priority given to intersections). If the iModulon consisted of genes with annotated co-regulation by 4 or more genes, a specific enrichment calculation was made to determine the enrichment statistics. If none of these adjustments yielded a significant enrichment, the iModulon was annotated as non-regulatory. All parameters and statistics related to calculation of TRN enrichments for regulatory iModulons are recorded in the iModulon metadata table, available in the GitHub repository. If any significant regulatory enrichments were found after applying this procedure, the iModulon was annotated as Regulatory and named according to the ruleset defined below in Case 1. Otherwise, the iModulon was assigned one of 4 additional categories (Genomic, Biological, Single-Gene Dominant, Uncharacterized), detailed in Cases 2-5 below, respectively.

iModulons were named and annotated according to the following ruleset:

General Rule #1: iModulon names must be fewer than 15 characters Rule #2: iModulon names must be unique. If iModulons would otherwise have the same name, append "-1", "-2", etc., as needed to disambiguate. By default, order the suffixes by decreasing explained variance, unless another numbering is specifically preferred (e.g. aligning Crp-1 and Crp-2 with Crp binding site classes).

Case 1 - Regulatory The iModulon has a significant regulon enrichment chosen as described above: Rule #1: Name the iModulon after the primary function of the enriched regulon(s) (e.g. the iModulon enriched for the CdaR regulon is named "Sugar Diacid") Rule #2: If no clear primary function is available for the iModulon, name the iModulon directly after the enriched regulon (e.g. the iModulon enriched for the CpxR regulon is named "CpxR", as CpxR controls a diverse set of functions). Exception #1: if the enriched regulon corresponds to a well-known global regulator (i.e. Fur, CRP, RpoS), name the iModulon after that regulator. Exception #2: if the name per Rule #1 would violate General Rule #1, name the iModulon directly after the enriched regulon (e.g. the iModulon enriched for the union of the FucR and ExuR regulons is named "FucR/ExuR" instead of "Fucose/Galacturonate/Glucuronate") Exception #3: if applying Rule #2, and the regulon enrichment involves an intersection between a global regulator and a local regulator (i.e. cooperative regulation), the global regulator may be dropped from the name (e.g. "NtrC-1" instead of "RpoN+NtrC-1", as RpoN is a larger-regulon sigma factor which co-regulates with the more-specific NtrC).

Case 2 - Genomic The iModulon activity profile has a clear correlation with a sample involving a specific genetic or genomic intervention: Rule #1: if the iModulon captures intentional knockout of a gene (e.g. geneA is knocked out in sampleA, and the iModulon has a large positive gene weight for geneA and a large negative activity level for sampleA, accounting for the lack of geneA expression in sampleA), name the iModulon "[gene name] KO" (e.g. baeR KO) Rule #2: Similarly, if the iModulon captures intentional overexpression of a particular gene, name the iModulon "[gene name] OE" (e.g. "malE OE") Rule #3: if the iModulon captures expression changes in relation to evolved samples (ALE), as determined by comparing the iModulon activities to known ALE samples, name the iModulon "[name of ALE project] Del" (for deletions),

"[name of ALE project] Amp" (for amplifications), or "name of ALE project] Mut" (for mixed effect mutations) (e.g. ROS TALE Del-1) Rule #4: if the iModulon also has a significant regulon enrichment as described above, prioritize the specific genetic/genomic change.

Case 3 - Biological The iModulon does not have a significant regulon enrichment, does not relate to a specific genetic or genomic change, but the member genes share a clear biological function: Rule #1: Name the iModulon after the shared biological function (e.g. the "LPS" iModulon consists of many genes related to lipopolysaccharide biosynthesis and export, though no significant regulon enrichment was found for this iModulon's genes).

Case 4 - Single-Gene Dominant The iModulon contains one specific gene with a gene weight at least twice as large as the next closest gene, does not fall into Case 2 - Genomic, and contains only the one highly-weighted genes, or at most 5 other genes with gene weights very close to the iModulon's threshold Rule #1: Name the iModulon after the dominant gene (e.g. the "ymdG" iModulon consists solely of the ymdG gene)

Case 5 - Uncharacterized The iModulon does not meet any of the previous criteria for naming Rule #1: Name the iModulon "UC-#" (short for "Uncharacterized"), with the number incrementing for each uncharacterized iModulon.

Differential iModulon activity computation

Differentially iModulon activities (DiMAs) were computed with a similar process as previously detailed [45]. For each iModulon, the average activity of the iModulon between biological replicates, if available, was computed. Then, the absolute value of the difference in iModulon activities between the two conditions was compared to the fitted log-normal distribution of all differences in activity for the iModulon. iModulons that had an absolute value of activity greater

than 5, and an FDR below 0.05 were considered to be significant. The number of DiMAs was computed for each unique pair of conditions within each project in the PRECISE-1K compendium, mirroring DEG computation.

Compiling the public K-12 Dataset

Data was compiled from NCBI SRA as described previously [160]. Initially, all data annotated as RNA-seq for E. coli was inspected. RNA-seq samples were discarded if the strain was not from a K-12 strain, if the strain was missing, or if the type of experiment was not actually RNA-seq. After initial curation, 3,125 samples remained. Next, these data were processed and quality controlled as described previously. 74% of samples (2,312) passed the RNA-seq quality control checks (FastQC, minimum reads mapped to coding sequences, non-outlier clustering). 58% of the original samples (1,816) had sufficient metadata annotation to verify biological replicates. Only conditions with at least 2 biological replicates were kept at this step. Finally, the 0.95 minimum replicate correlation threshold was applied, yielding the final set of 1,675 high-quality publicly-available samples (54% of the original set). Next, these 1,675 samples were combined with the 1,035 samples of PRECISE-1K to yield the "Public K-12" dataset, comprising 2,710 curated, high quality expression profiles for Escherichia coli strain K-12. The log_2[TPM], raw read count, QC data files, and sample metadata for the high-quality public samples may be found in the data directory of this project's GitHub repository. After centering the Public K-12 dataset to the PRECISE-1K control condition, iModulons were computed and annotated in the same manner as described above.

Acknowledgements

This research was supported by the Novo Nordisk Fonden (NNF20CC0035580).

Chapter 4 in part is a reprint of material published in:

- **CR Lamoureux**, KT Decker, AV Sastry, K Rychel, Y Gao, JL McConn, DC Zielinski, and BO Palsson. 2023. "A multi-scale expression and regulation knowledge base for *Escherichia coli*." *Nucleic acids research*, gkad750. The dissertation author was the primary author.

Chapter 5

Conclusions

The study of biology now more than ever necessitates big data analytics. Experimental methods and knowledge have advanced to the point that the study of life at the systems level is possible. In particular, large-scale nucleic acid sequencing data enables study of the flow of information within a single genome, across multiple genomes, and from genome to transcriptome. The study of these datasets promises to yield significant insight into the complexity of living cells. In this thesis, we introduce a trio of data analysis frameworks that convert big data to biological knowledge at multiple scales.

In the introduction, we described how the availability of large-scale DNA and RNA sequencing data has revealed multiple scales of information encoding and activation. In particular, we highlighted *Escherichia coli* as a model bacterium for which significant genome annotation, genome sequence, and transcriptome data is present. We have since developed analytical frameworks that scalably provide biological insight from this data.

In Chapter 2 we investigated genome annotation information by constructing the Bitome, a formalized and binarized representation of genomic information and single base pair resolution.

We identified unequal patterning of annotation information and leveraged the information for classification of adaptively-mutated genes and quantitative prediction of mRNA transcript levels. Importantly, we demonstrated extensibility of the framework to other strains and organisms and disseminated the relevant software tools.

Chapter 3 expanded on the Bitome, extending its annotation across over two thousand *E. coli* strains to assess non-coding sequence variation in promoter and 5' UTR regions. This "alleleome" revealed significant conservation in functional non-coding features, identified variability in transcription factor binding site conservation, and contrasted wild-type non-coding variation with adaptive mutations. This work also revealed the sufficiency of non-coding alleles to discriminate phylogroups. Again, the underlying analytical framework and software were published to facilitate similar analysis of other organisms.

Finally, Chapter 4 shifted attention to the regulatory patterns in transcriptomic data. We assembled the largest single-protocol RNA-seq compendium for *E. coli* - PRECISE-1K - and described genome-scale expression trends. Centrally, we used unsupervised machine learning to extract 201 regulatory signals (iModulons) from PRECISE-1K that capture the majority of the known transcriptional regulatory network. These iModulons enable systems-level analysis of regulatory information under different environmental conditions. Critically, the dataset and associated software packages have been published to facilitate analysis of new data against PRECISE-1K as a large-scale control. Nearly two-thousand publicly-available transcriptomes were combined with PRECISE-1K and decomposed to yield a similar iModulon structure, highlighting the method's scalability and robustness.

While these aims have advanced our understanding of the patterning, variation, and activation of genomic information, additional goals remain. Interoperability of analysis methods

is an important component of systems biology; a desirable next step is to more explicitly combine these different-scale analyses. Integrating additional data type and methods - such as proteomics and metabolic modeling, respectively, could further assist in unraveling the complex relationship between genotype and phenotype.

Overall, this thesis presents important advancements in the conversion of big biological data to systems-level, actionable knowledge.

Appendix A

The Bitome: Digitized genomic features reveal fundamental genome organization - Supplementary Information

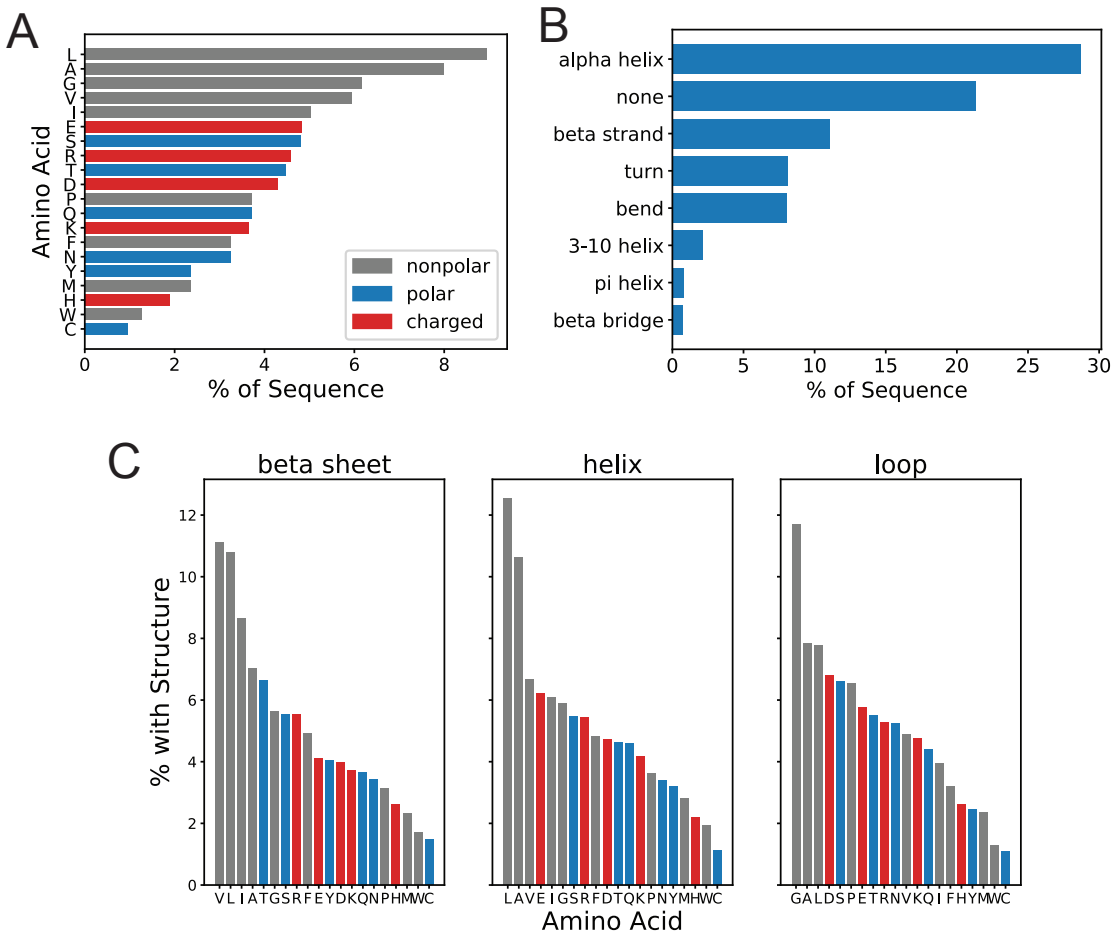


Figure A.1: Amino acid and secondary structure information are easily cross-referenced. **A)** Percentages of the total genomic positions coding for each of the amino acids. This calculation does not double-count genomic positions that code for the same amino acid on the forward and reverse strands. **B)** Percentages of total genomic positions devoted to coding for different predicted secondary structures. Predictions from the DSSP algorithm (see Methods). **C)** Amino acid preferences of secondary structural groups. See legend from **A** for color code. Beta sheet: beta bridge or beta strand; helix: alpha helix, 3-10 helix or pi helix; loop: turn or none.

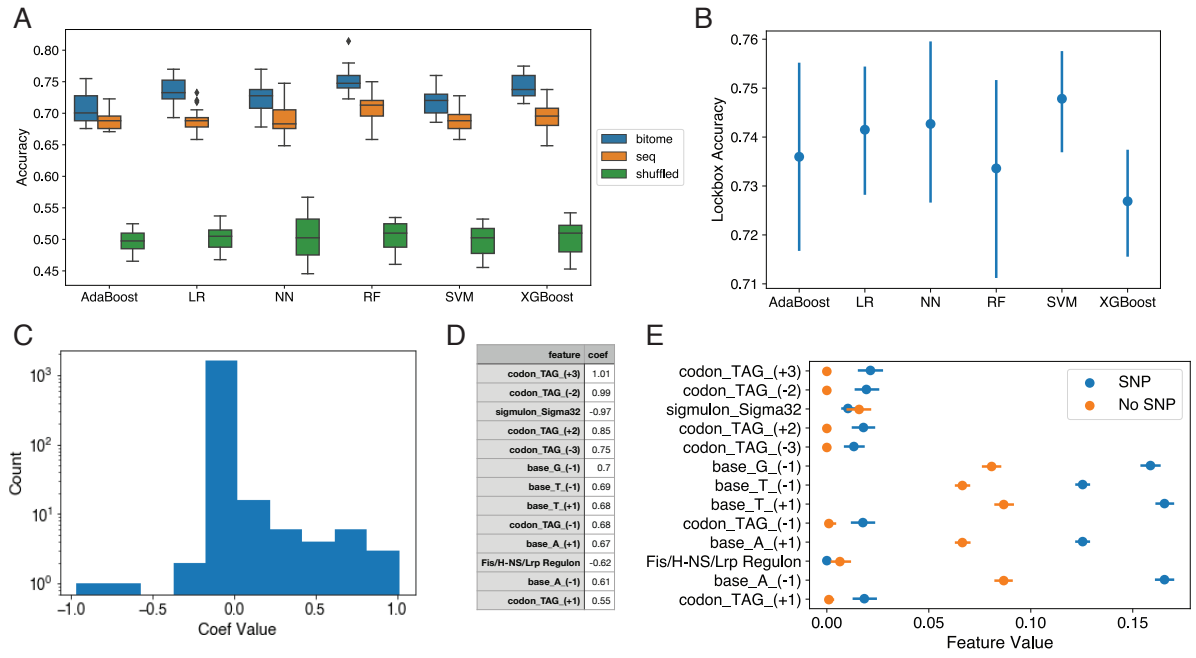
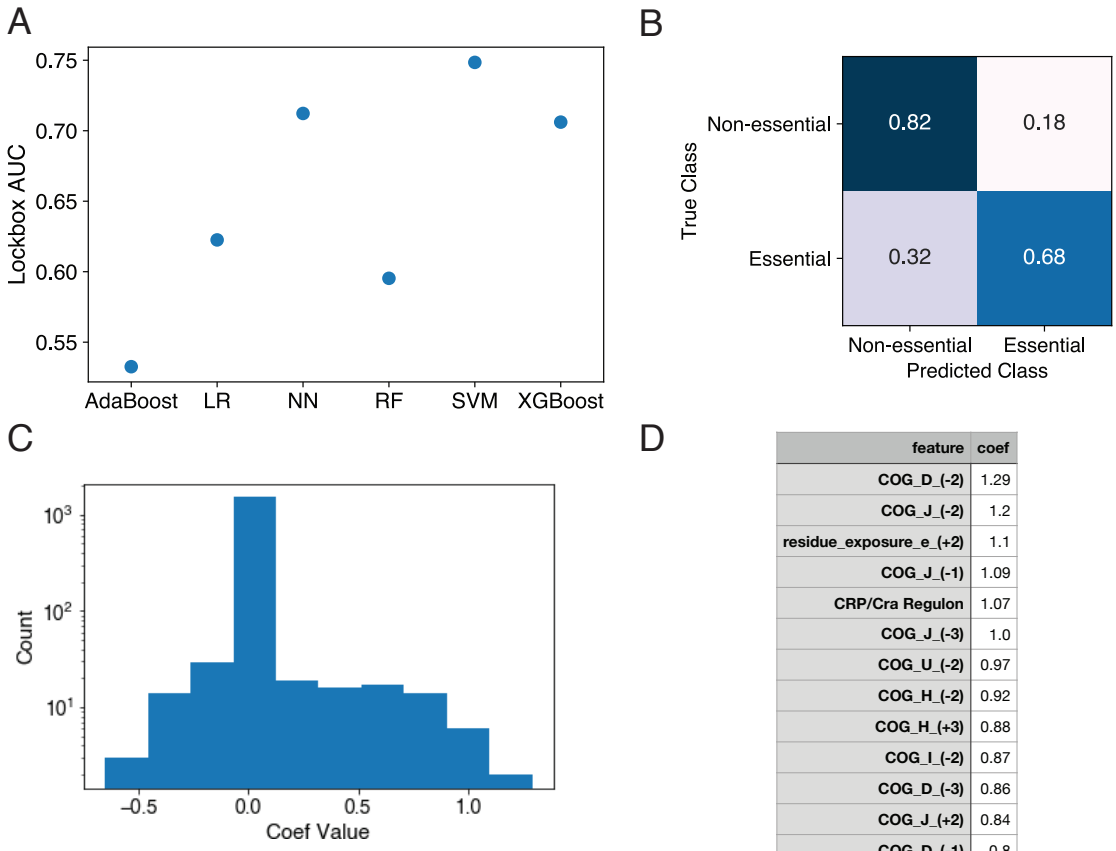


Figure A.2: Support vector machine classifier selects Bitome features to predict genes with ALE SNPs. **A)** Performances of 6 different out-of-the-box models at classifying genes with ALE SNPs. Bitome = used all bitome features; seq = sequence only; shuffled = shuffled target labels. n=25 for each group, composed of 5-fold cross validation for each of 5 downsamplings. AdaBoost = adaptive boosted tree; LR = logistic regression; NN = neural network; RF = random forest; SVM = support vector machine; XGBoost = extreme gradient-boosted trees. **B)** Final performance of models on held-out, lockbox test data. Points are mean of performance on 5 downsamplings, bars are standard deviation. **C)** Feature importances of final support vector machine classifier. **D)** Table of Bitome features identified as important for classification. Positive coefficients indicate importance for predicting the SNP class, and negative coefficients indicate importance for predicting the No SNP class. **E)** Class comparison of values (after min-max scaling) of the important features from panel D in the training set. Points are mean, bars are standard deviation. No SNP n=1010, SNP n=2339.



feature	coef
COG_D_(-2)	1.29
COG_J_(-2)	1.2
residue_exposure_e_(+2)	1.1
COG_J_(-1)	1.09
CRP/Cra Regulon	1.07
COG_J_(-3)	1.0
COG_U_(-2)	0.97
COG_H_(-2)	0.92
COG_H_(+3)	0.88
COG_I_(-2)	0.87
COG_D_(-3)	0.86
COG_J_(+2)	0.84
COG_D_(-1)	0.8
COG_I_(+1)	0.8

Figure A.3: Classification of essential genes. **A)** Final performances of models on held-out, lockbox test data. AUC = area under the receiver operating characteristic curve. **B)** Confusion matrix for support vector machine classifier. Scores are accuracy, normalized to true class. n=838 in held-out, lockbox test set. **C)** Feature importances of final support vector machine classifier. **D)** Table of Bitome features identified as important for predicting essential genes. COG = cluster of orthologous groups; D = cell cycle control, cell division, chromosome partitioning; J = translation, ribosomal structure and biogenesis; U = intracellular trafficking, secretion, and vesicular transport; H = coenzyme transport and metabolism; I = lipid transport and metabolism.

Appendix B

Escherichia coli functional non-coding regions are highly conserved - Supplementary Information

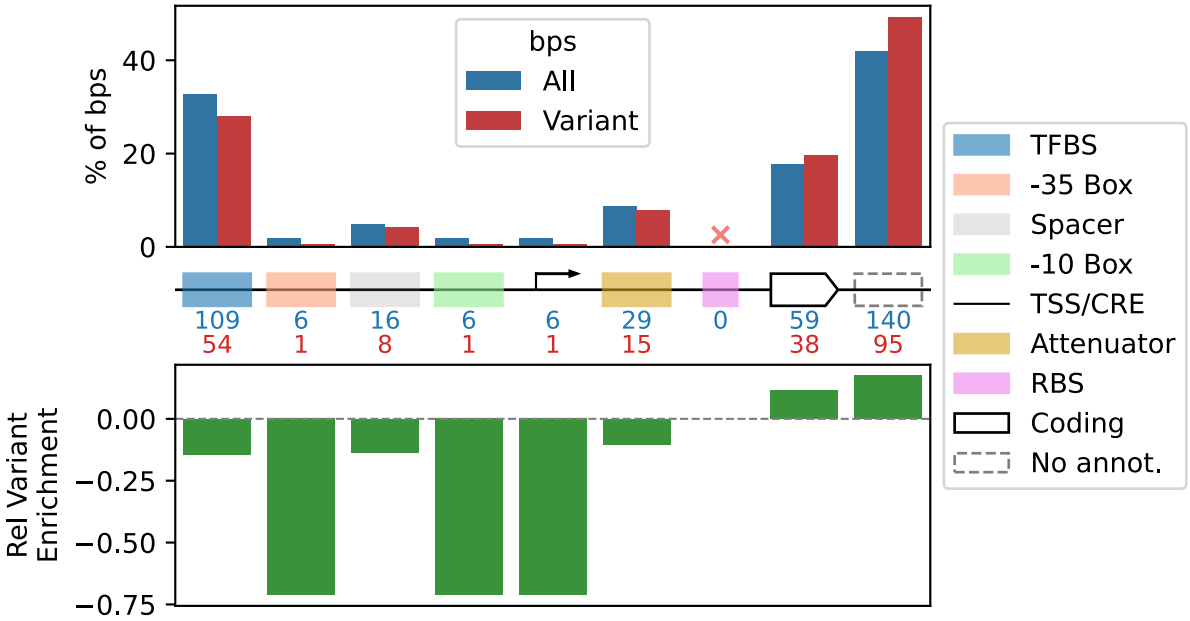


Figure B.1: Breakdown of base pairs in *aceB* region based on presence in annotated promoter features. Blue bars indicate % of all 331 base pairs in the region that are annotated with each category. Red bars indicate % of 193 variant base pairs belonging to each category. Note that due to overlaps, percentages add up to larger than 100%. Blue and red numbers below the schematic indicate numbers of base pairs in each category. Red X in the bar plot indicates features with no presence in this region. Green bars indicate relative enrichment of variant base pairs in each category: (% variant bps in category / % all bps in category) / % all bps in category. E.g., 109/331 (33%) of all bps are in a TF binding site, along with 54/193 (28%) of variant bps: (28% - 33%) / 33% = -0.15.

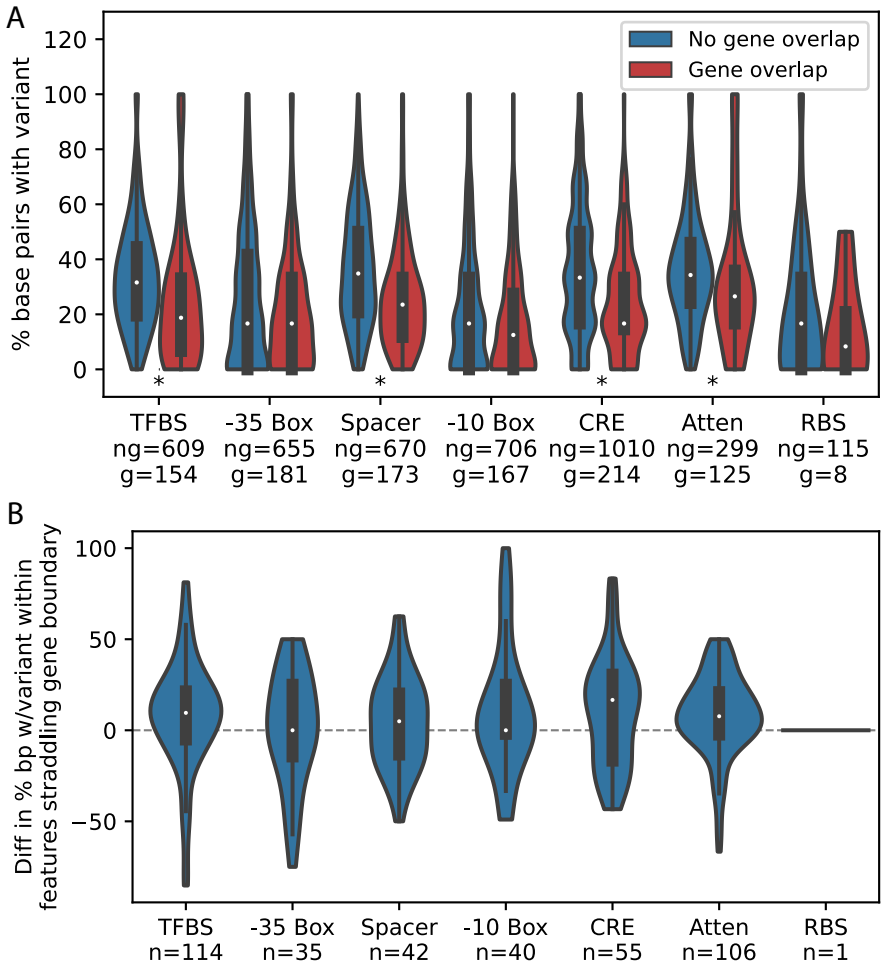


Figure B.2: Overlap with coding regions influences variation in non-coding features. **A)** Violin plots comparing distributions of percentages of base pairs with a variant within non-regulatory features across the alleleome. For each feature type, the blue violin plot shows the variant base pair percentage distribution for instances of that feature type that do not overlap with a coding gene in the reference strain (on either strand), while the red violin plot shows the opposite. ng = # of instances of feature type with no gene overlap; g = # of instances w/gene overlap. Asterisks indicate statistically significant difference between groups within feature type (Mann-Whitney U, FDR controlled at 0.01). **A)** Distributions of differences in variant base pair percentage between coding-overlap and non-coding overlap portions of non-coding features that straddle gene boundaries. n = # of instances of feature type that straddle coding gene boundary. Each value in each distribution is computed by finding the % of base pairs in the non-coding sub-portion of a non-coding feature that have a variant and subtracting the % of variant base pairs in the coding sub-portion. E.g. y-values above 0 indicate more variants in the non-coding portion of the feature.

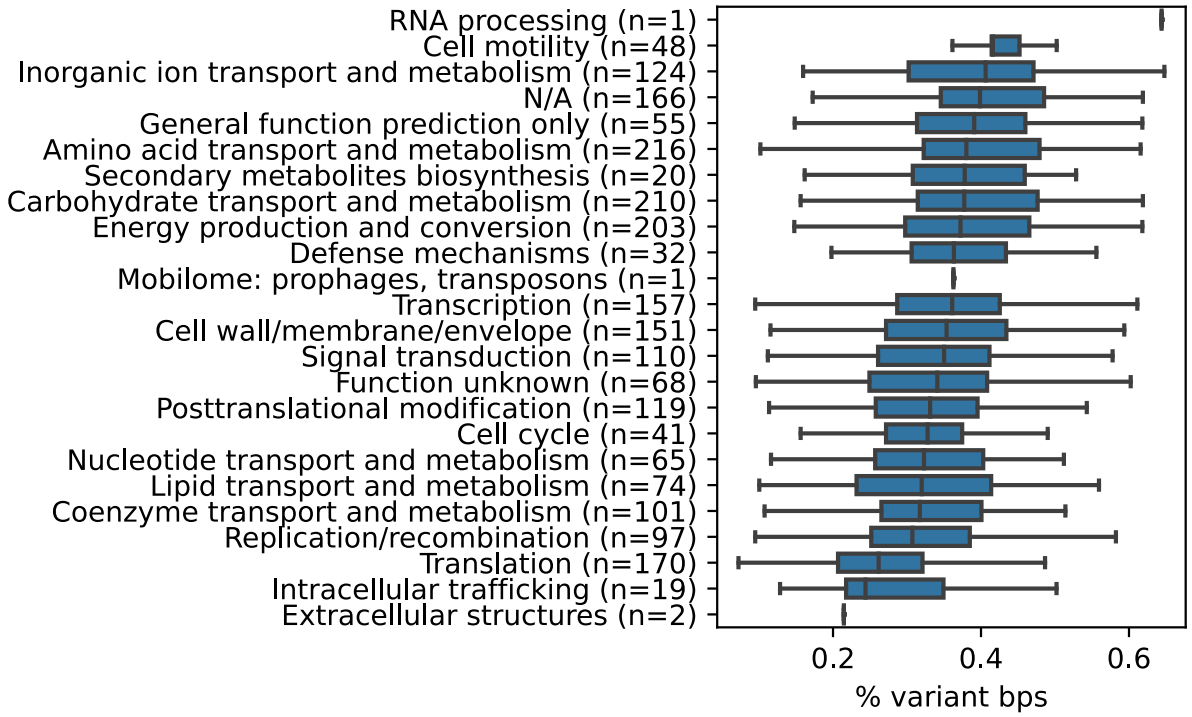


Figure B.3: Distributions of percentage of variant base pairs in regions transcribing different COGs.

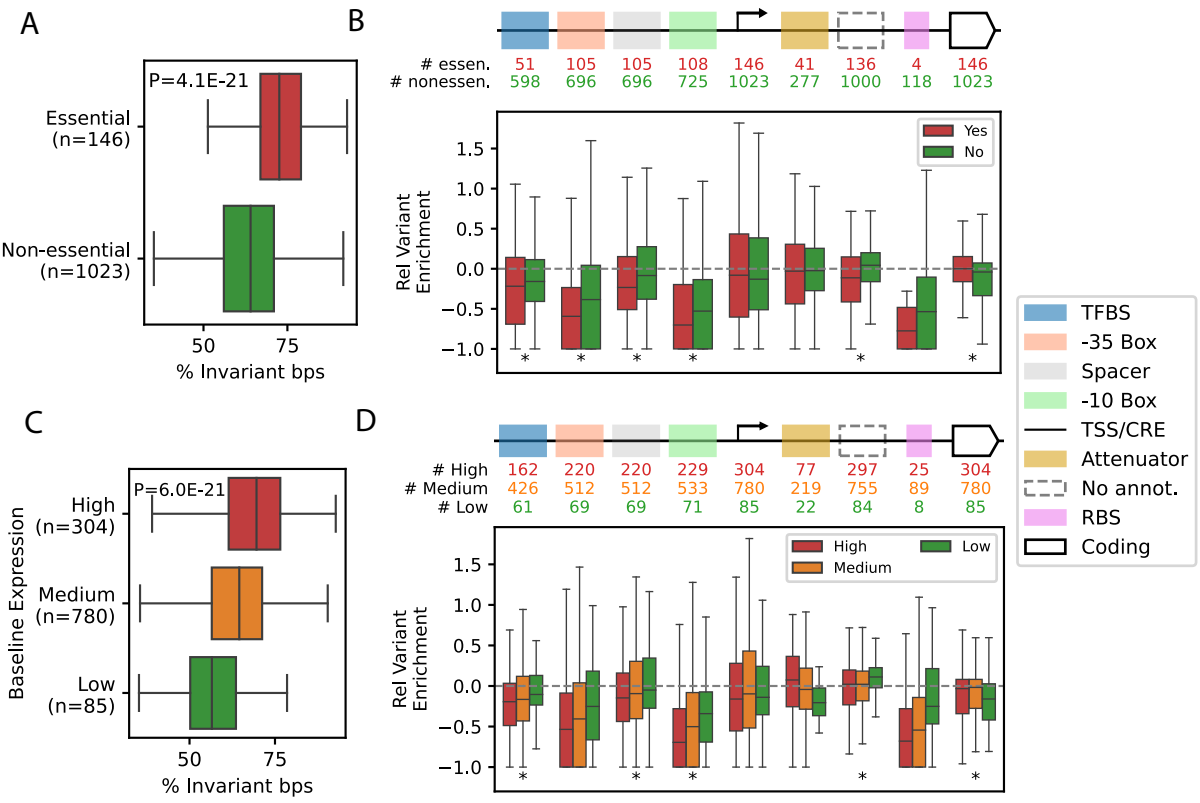


Figure B.4: Non-coding conservation is related to phenotypic outcomes. A) Boxplot comparing % invariant base pairs in regions encoding at least one essential gene (Yes) against regions with no essential genes transcribed (No). Essentiality defined per the Keio collection [53] P value from Mann-Whitney U test. B) Relative variant enrichment (degree of conservation relative to sequence length) for non-coding feature categories, segmented by essentiality. Asterisks indicate features with significant difference in variant enrichment between essential and non-essential regions (Mann-Whitney U test, FDR controlled at 0.1). C) Boxplot comparing % invariant base pairs in regions with different baseline expression categories as defined by PRECISE-1K knowledge base [54]. Expression for a non-coding region computed as median of expression levels of genes transcribed from region. P value from one-way ANOVA. D) Relative variant enrichment for three baseline expression categories. Asterisks indicate features with significant differences among groups (one-way ANOVA, FDR controlled at 0.1).

Appendix C

A multi-scale *Escherichia coli* expression and regulation knowledge base - Supplementary Information

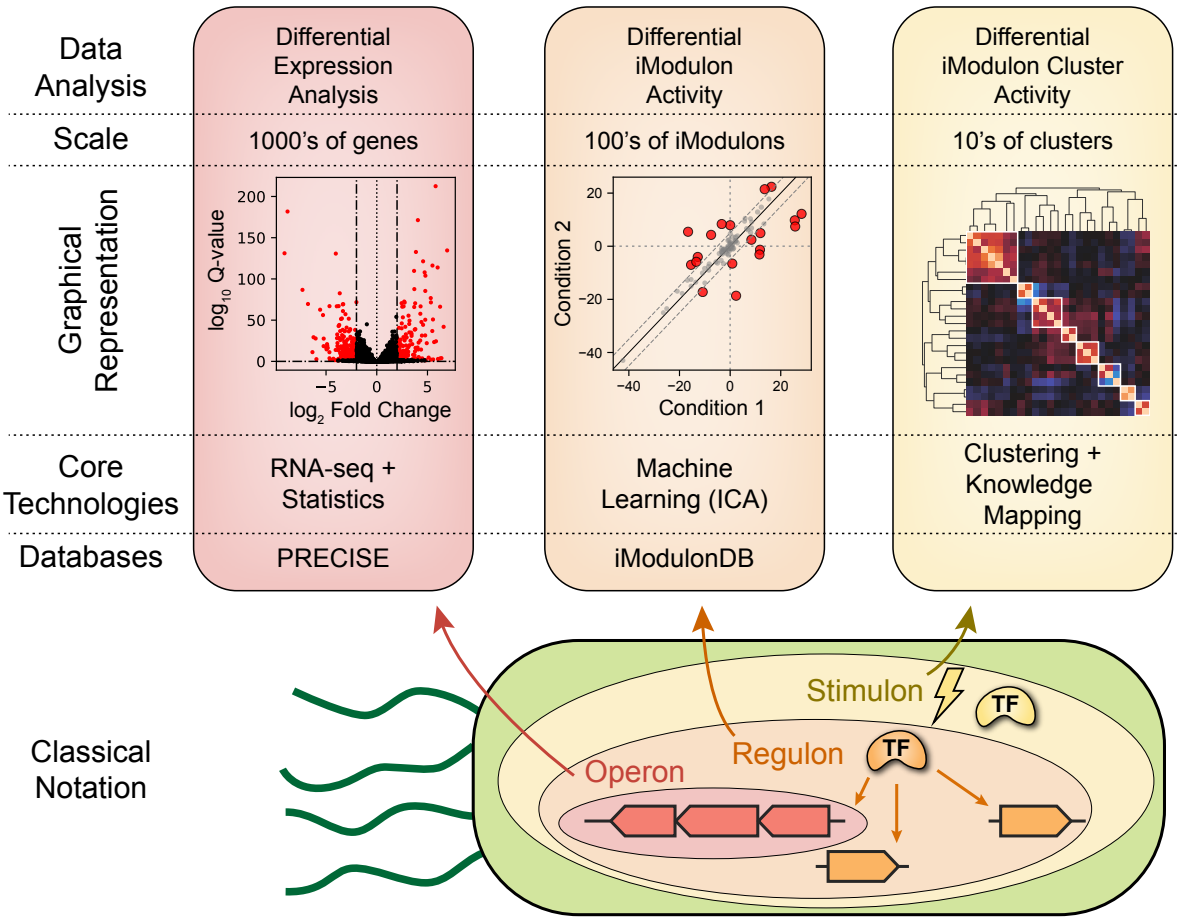


Figure C.1: Multi-scale analysis of PRECISE-1K. The levels of analysis approximately correspond to the definition of an operon and a regulon, and also a quantitative definition of the notion of a stimulon. The 'scale' indicates the reduction of dimensionality over the levels shown.

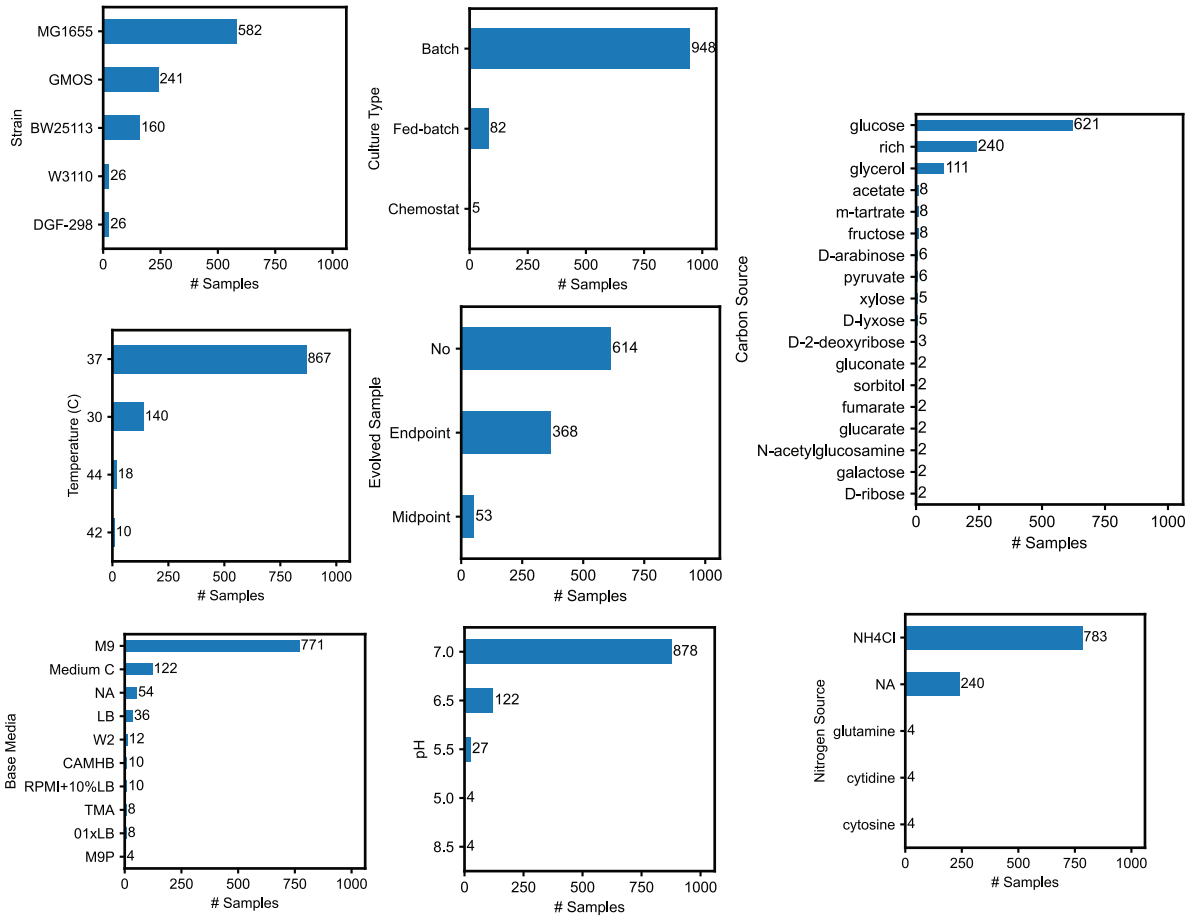


Figure C.2: Breakdown of major growth conditions for PRECISE-1K.

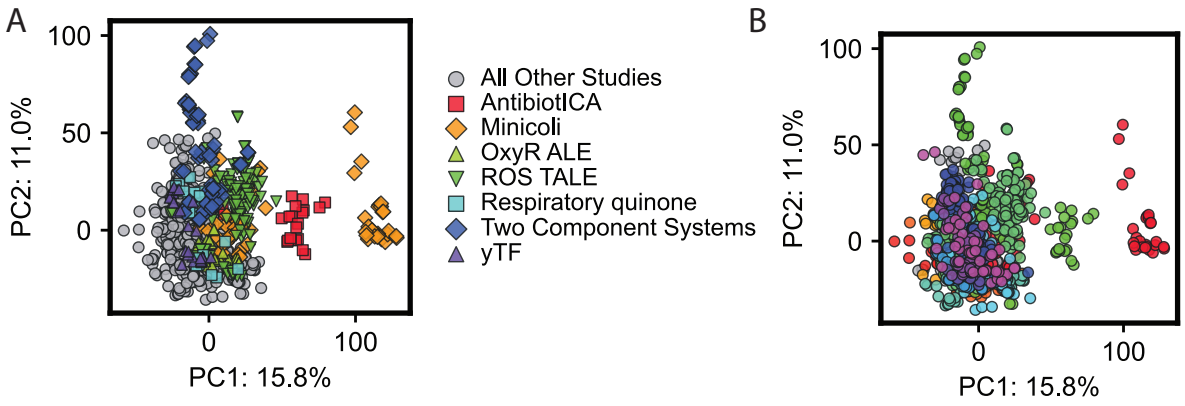


Figure C.3: Principal component analysis (PCA) of PRECISE-1K. **A)** First 2 principal components, colored by project (n=1035 samples). **B)** First 2 principal components, colored by each of 21 distinct RNA-seq library preparers.

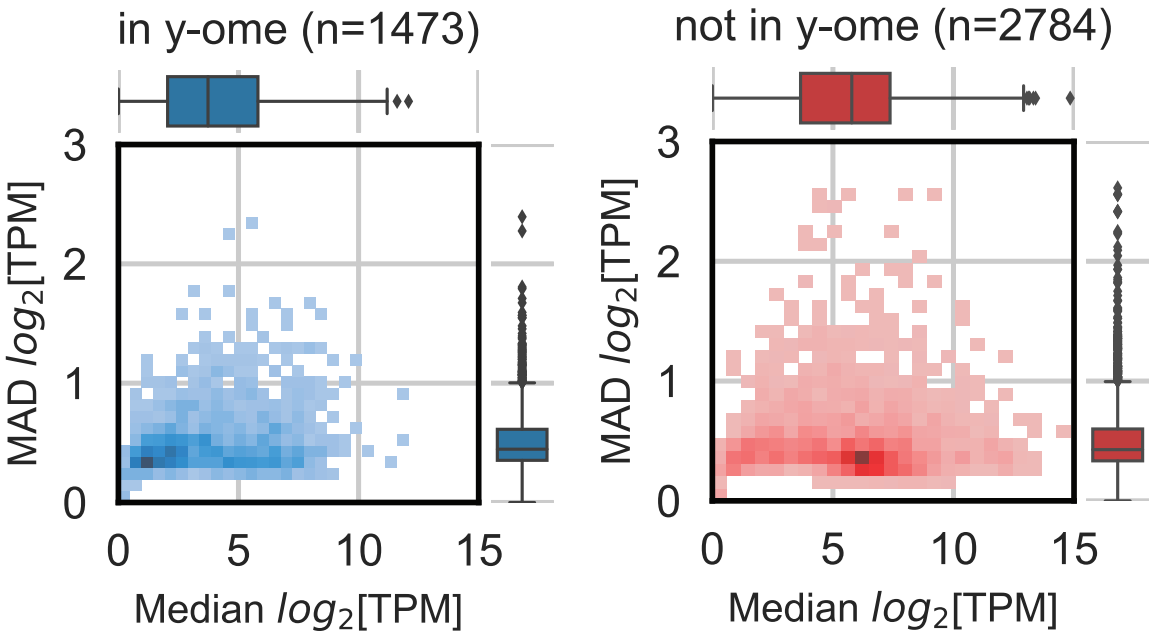


Figure C.4: Median vs MAD expression 2-D histogram, separated by annotation status [49]. Blue = y-ome (poorly-annotated genes); red = well annotated genes.

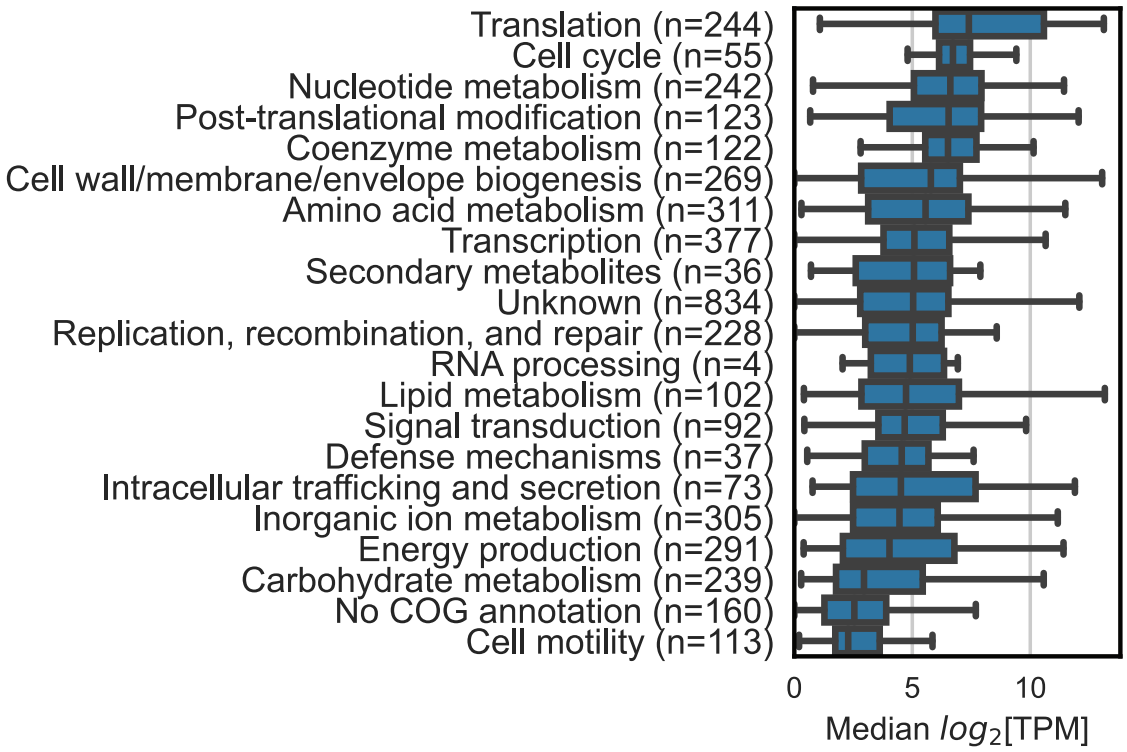


Figure C.5: Breakdown of gene expression by COG category across PRECISE-1K.

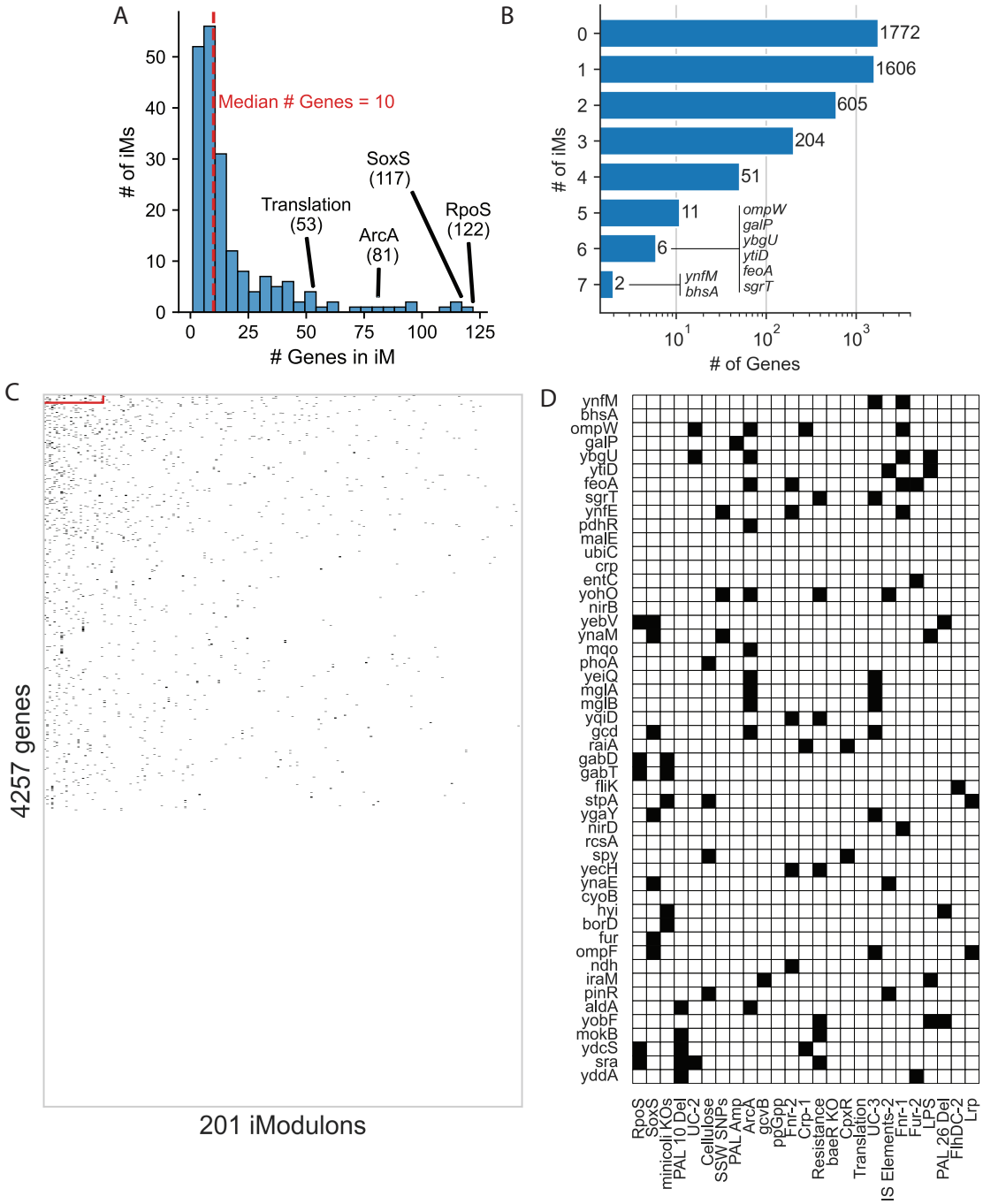


Figure C.6: iModulon gene membership breakdown. **A)** Histogram of iModulon sizes. n=201 iModulons. **B)** Breakdown of genes by number of iModulons of which they are a member. n=4257 genes. 2485 genes are members of at least 1 iModulon. **C)** Representation of binarized M matrix, relating genes (rows) to iModulons (columns). Black cells indicate membership of that row's gene in that column's iModulon. Rows and columns are sorted in descending order by sum, from upper left to lower right. Red boxed region is displayed in panel **D**. **D)** Zoomed-in display of the 50-gene x 25-iModulon red boxed region from panel **C)**.

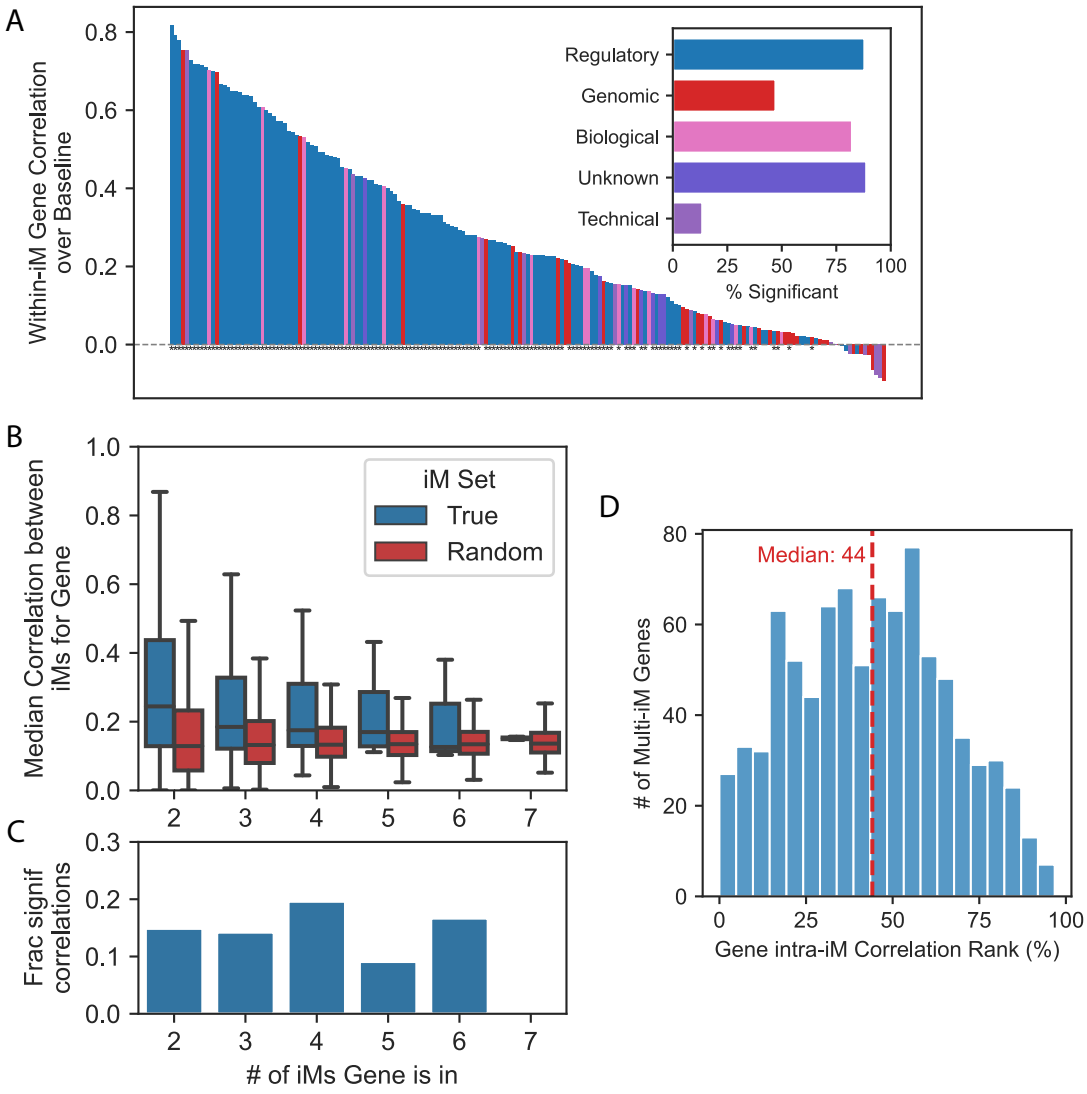


Figure C.7: Within-iModulon gene correlations and multi-gene iModulon analysis. **A)** Difference between median of pairwise gene correlations in each iM and expectation. Expectation from bootstrapping: sample 1000 iM-size gene groups, compute median pairwise r, take overall median of medians. * indicates significance (against bootstrapped null, FDR at 0.05). Inset: percentage of iMs with significant pairwise correlations within their genes. **B)** Median correlation between iMs for 879 multi-iM genes, by number of iMs for gene. For each multi-iM gene, median of pairwise correlations between that gene's iMs was computed. That value is one value in blue box plot for appropriate number of gene iMs. Random (red) box plots indicate distribution of pairwise correlation medians for 1000 randomly-selected groups of iMs. **C)** Fraction of multi-iM genes significantly correlated iMs. Expectation from panel B. **D)** Histogram of "rank percentile" of each of 879 multi-iM gene's intra-iM correlation as compared to other iM genes. E.g. of 9 genes in the Salicylic Acid iM, multi-iM gene bhsA has the highest median correlation. It's ranked 1 out of 9 by intra-iM correlation, thus a "rank percentile" of 88.9%.

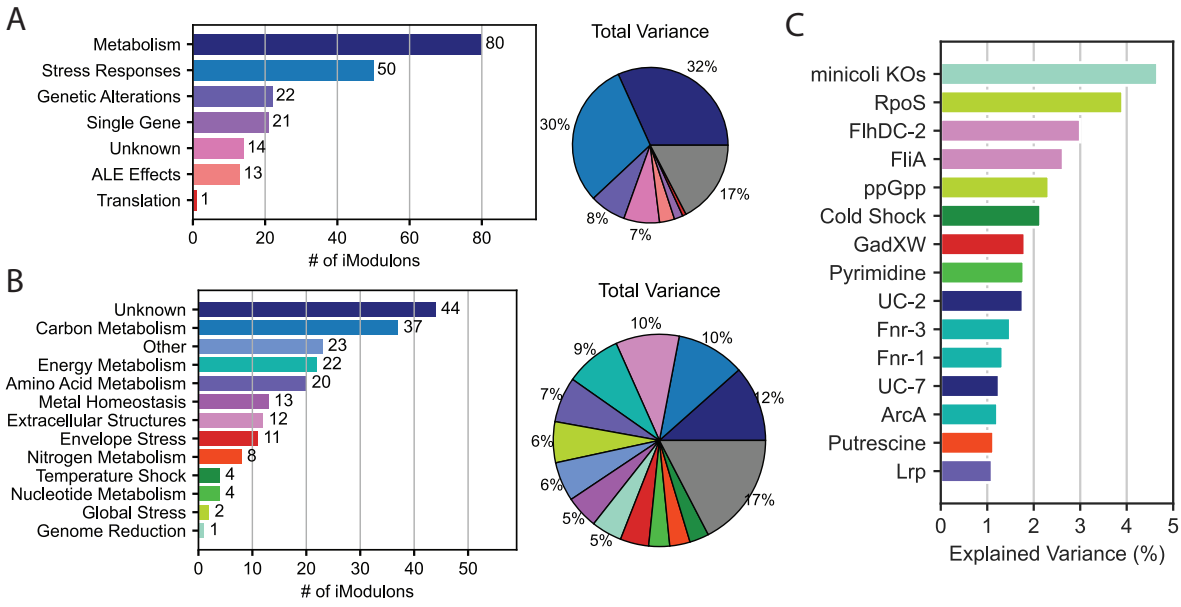


Figure C.8: Alternate iModulon categorizations and high-variance iModulons. **A)** Breakdown of iModulons based on system annotation. ALE = adaptive laboratory evolution. **B)** Breakdown of iModulons based on functional annotation. **C)** Top 15 iModulons by explained variance. Color code same as panel **B**.

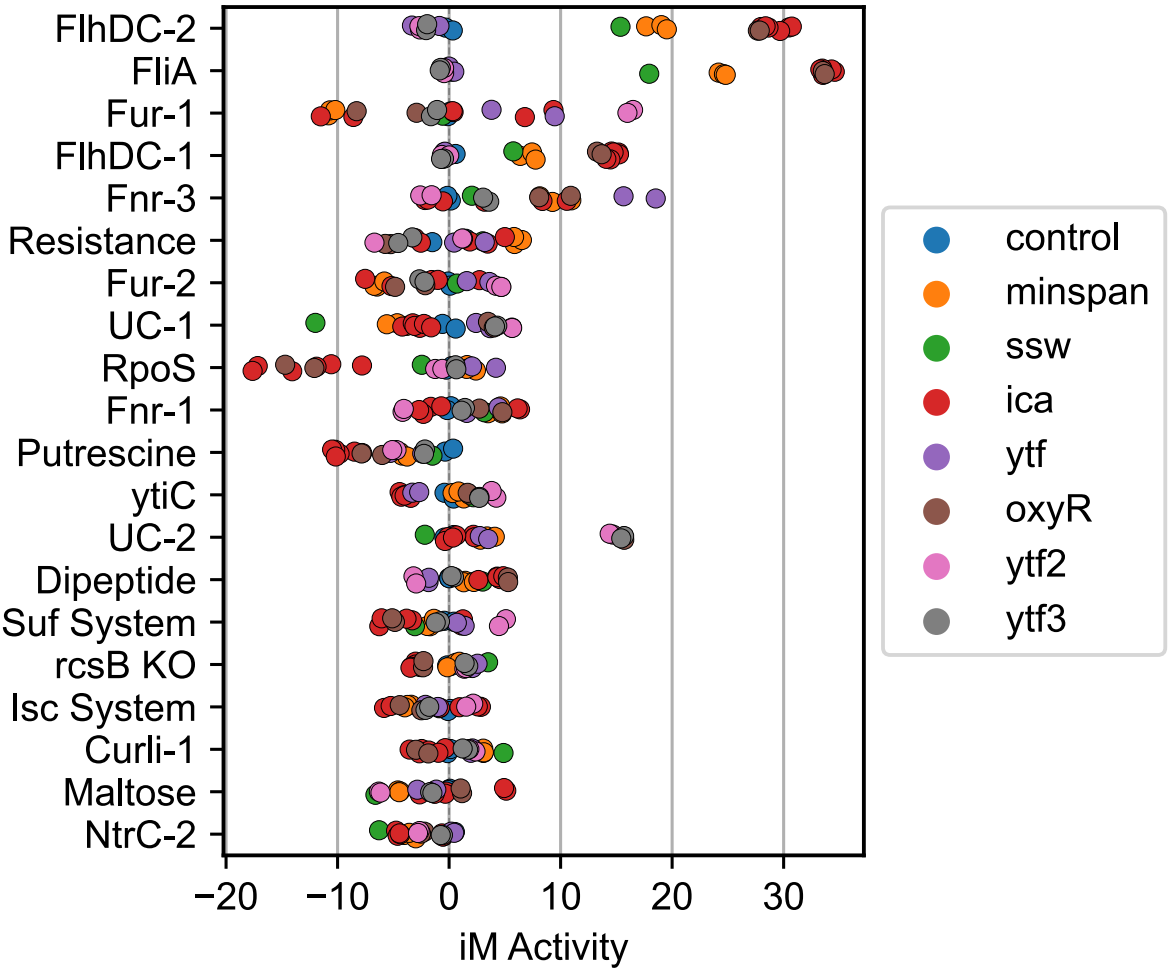


Figure C.9: Most variant iModulon activities in control conditions across projects. iModulon activities with the highest median absolute deviation across 20 samples of wild-type growth in M9 medium with glucose across 8 projects are displayed.

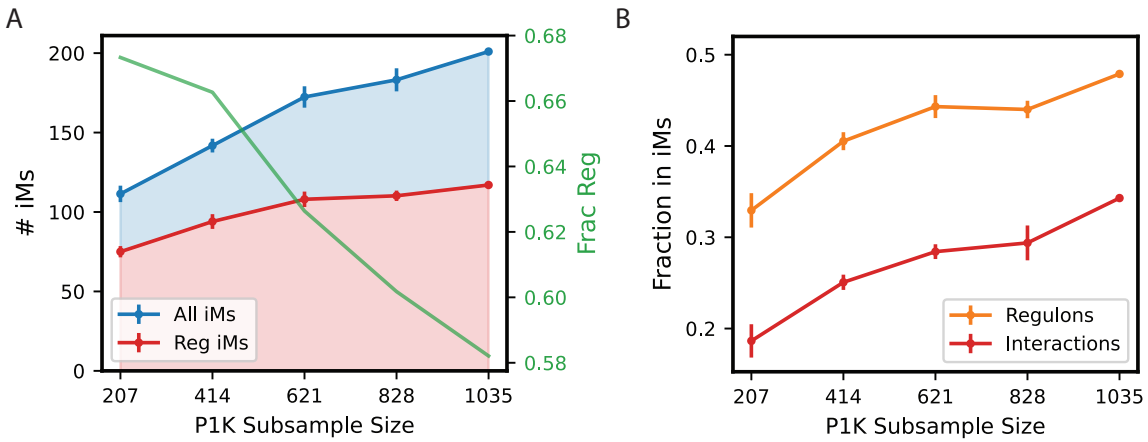


Figure C.10: PRECISE-1K subsample regulatory coverages. **A)** Numbers of iModulons extracted from each random subsample size (largest subsample size indicates PRECISE-1K iModulons themselves). Points are an average of 5 distinct trials; error bars indicate standard deviation. Frac reg = fraction of all iModulons that are regulatory (i.e. ratio of red value to blue value for each subsample size). Each subsample trial was run "additively"; i.e. for each of the 5 trials, 207 (20% of PRECISE-1K) random samples were chosen. Then, 207 more random samples from the unchosen set were added to the initial 207 to create the 414-scale subsample, and so on. **A)** RegulonDB enrichment fractions for regulons and regulatory interactions across all subsamples. Percentages correspond to all RegulonDB evidence levels (as in upper sets of bars in Fig. 4.2D). Regulatory interaction percentages refer to fraction of regulatory interactions directly captured by subsample (i.e. not "subsample+" as in Fig. 4.2D).

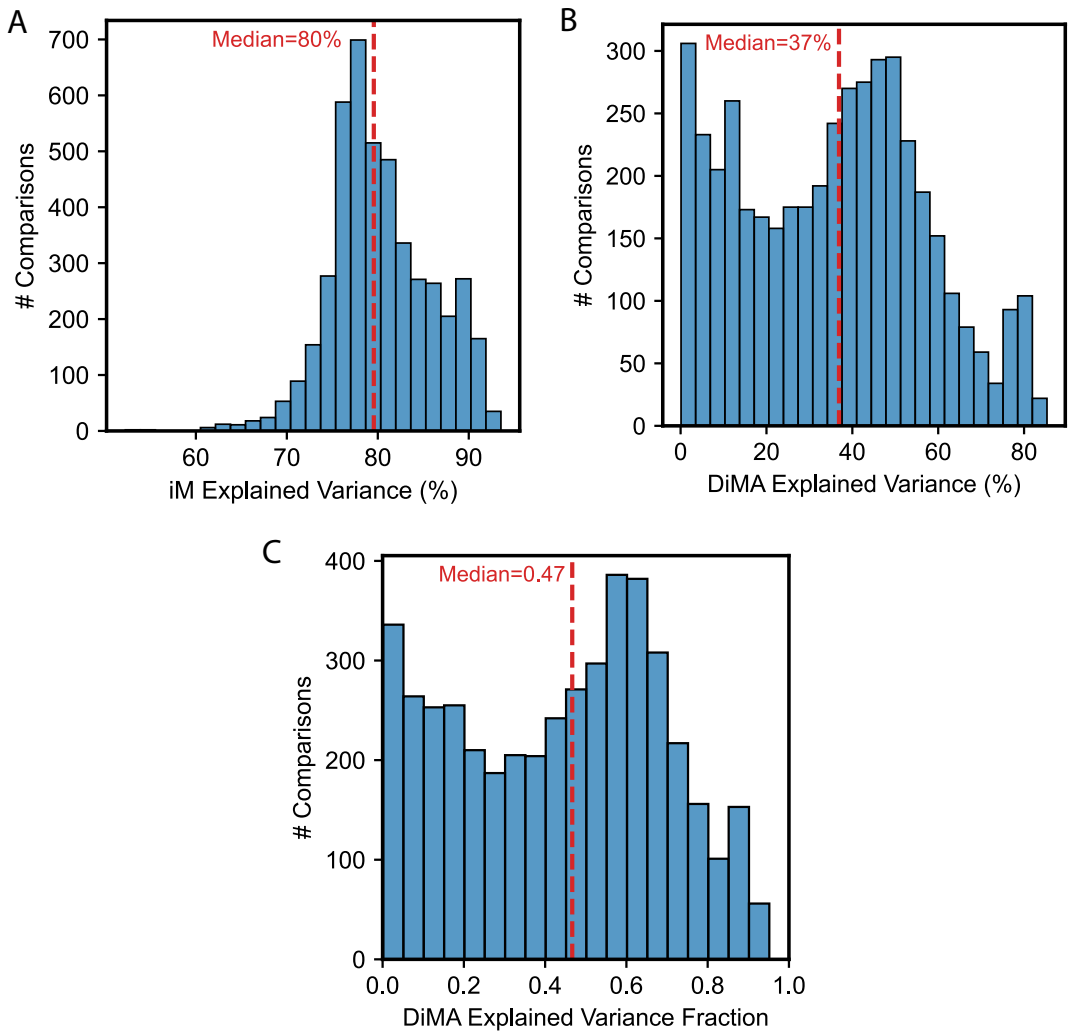


Figure C.11: DiMAs capture a variable amount of variance across condition comparisons. **A)** Histogram of percentage of variance explained by all iModulons for all condition comparisons with at least one DiMA (n=4483). **B)** Histogram of percentage of variance explained by iModulons with significant DiMA only. **C)** Histogram of fraction of all iModulon variance explained by iModulons with significant DiMA.

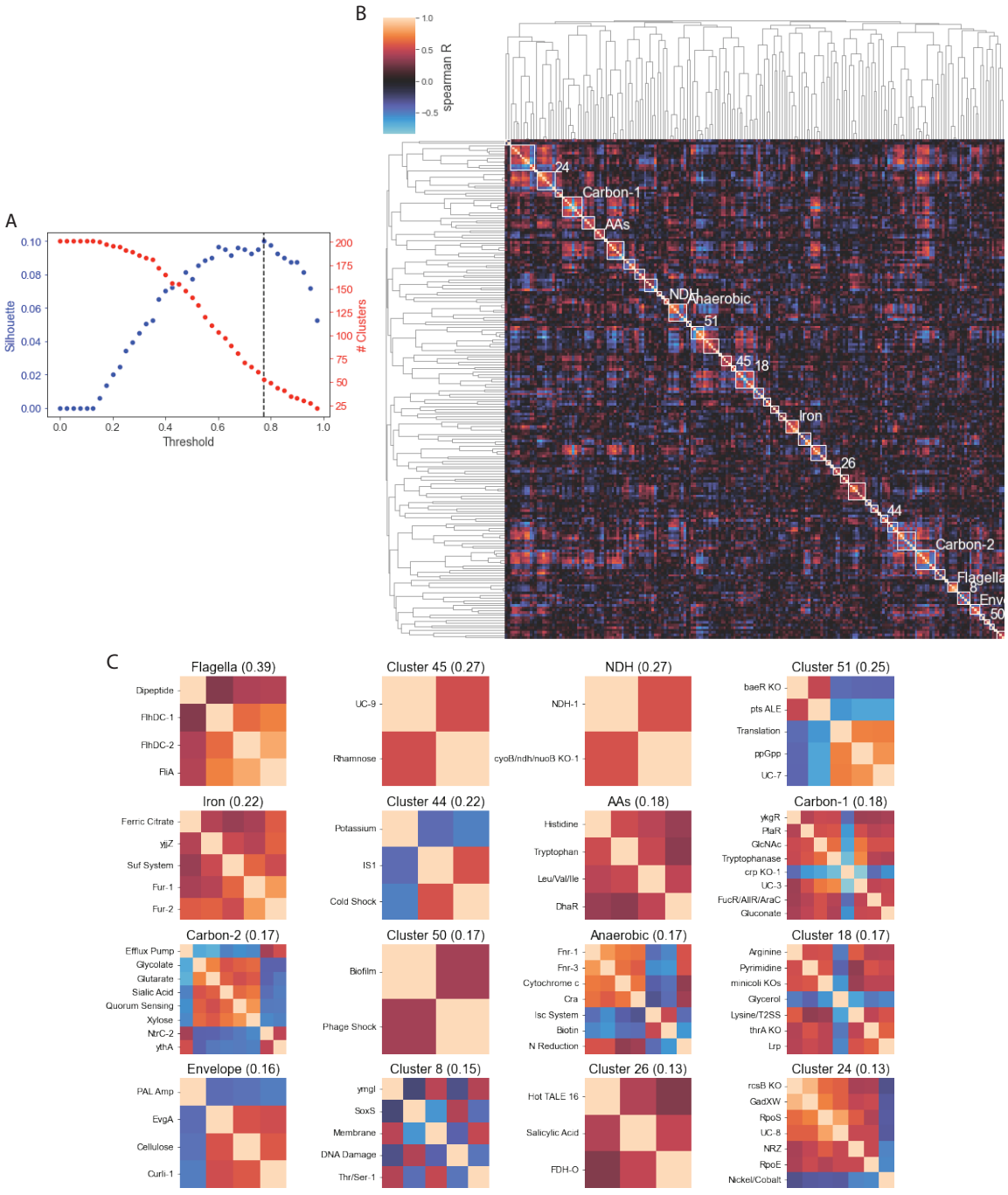


Figure C.12: iModulon activity clustering for PRECISE-1K: defining stimulons. **A)** Automatic distance threshold determination. **B)** Clustermap of PRECISE-1K iModulon activities. Colorbar indicates Spearman's r. 16 most distinct clusters labeled with yellow text corresponding to panel **C**. **C)** 16 most distinct clusters of iModulon activities from PRECISE-1K (silhouette score).

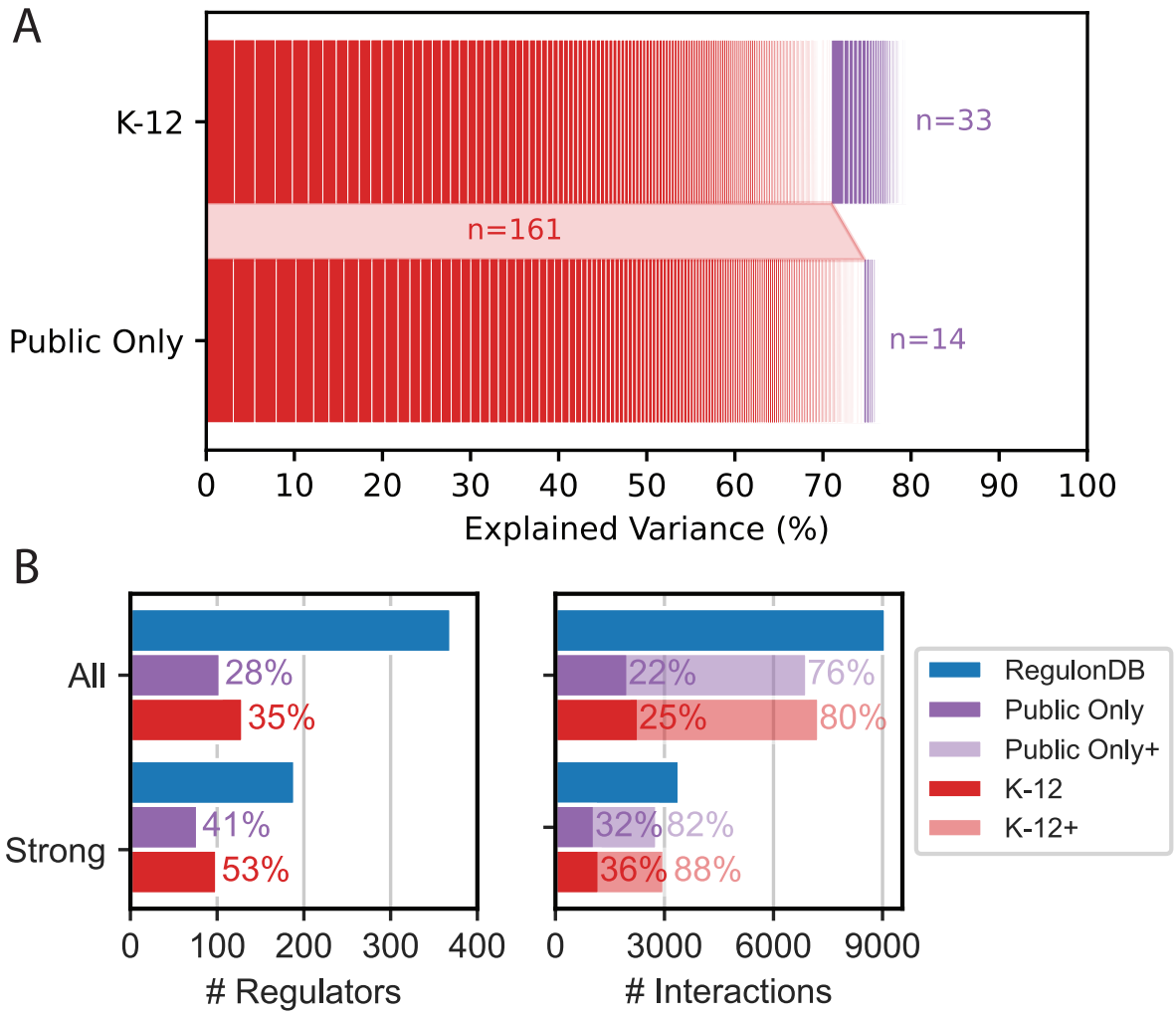


Figure C.13: Comparison of iModulons extracted from Public K-12 (i.e. public samples and PRECISE-1K) and from Public Only (1,675 public samples without PRECISE-1K). **A)** iModulon comparison between K-12 and Public Only. Same characteristics as Fig. 4.5D. Red = iModulon extracted from both datasets; purple = iModulon unique to dataset. **B)** Comparison of regulatory coverages for K-12 and Public Only. See Fig. 4.5C.

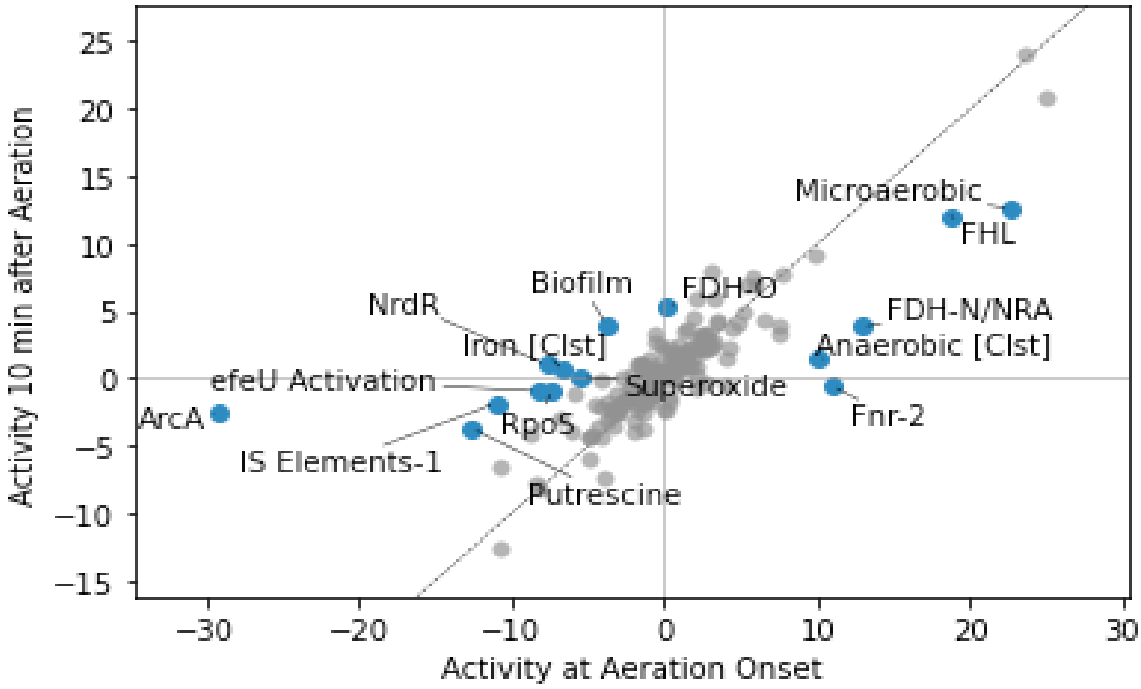


Figure C.14: DiMA plot between onset of aeration and 10 minutes post-aeration with activity clusters (stimulons) included (indicated with [Clst] suffix).

Bibliography

[1] Crick FH (1958) On protein synthesis. In: Symp Soc Exp Biol. volume 12, p. 8.

[2] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. science 269: 496–512.

[3] Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic escherichia coli. Nature reviews microbiology 2: 123–140.

[4] Nguyen Y, Sperandio V (2012) Enterohemorrhagic e. coli (ehec) pathogenesis. Frontiers in cellular and infection microbiology 2: 90.

[5] DuPont HL, Formal SB, Hornick RB, Snyder MJ, Libonati JP, Sheahan DG, LaBrec EH, Kalas JP (1971) Pathogenesis of escherichia coli diarrhea. New England Journal of Medicine 285: 1–9.

[6] Poirel L, Madec JY, Lupo A, Schink AK, Kieffer N, Nordmann P, Schwarz S (2018) Antimicrobial resistance in escherichia coli. Microbiology spectrum 6: 6–4.

[7] Karlowsky JA, Kelly LJ, Thornsberry C, Jones ME, Sahm DF (2002) Trends in antimicrobial resistance among urinary tract infection isolates of escherichia coli from female outpatients in the united states. Antimicrobial agents and chemotherapy 46: 2540–2545.

[8] 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. Nature communications 10: 935.

[9] Mohamed ET, Wang S, Lennen RM, Herrgård MJ, Simmons BA, Singer SW, Feist AM (2017) Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. Microbial cell factories 16: 1–15.

[10] Goeddel DV, Kleid DG, Bolivar F, Heyneker HL, Yansura DG, Crea R, Hirose T, Kraszewski A, Itakura K, Riggs AD (1979) Expression in escherichia coli of chemically synthesized genes for human insulin. Proceedings of the National Academy of Sciences 76: 106–110.

[11] Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with escherichia coli. Nature 461: 1243–1247.

[12] Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. Bioinformatics 24: 2044–2050.

[13] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ (2007) A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. Molecular systems biology 3: 121.

[14] Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, et al. (2023) Introducing the bacterial and viral bioinformatics resource center (bv-brc): a resource combining patric, ird and vipr. Nucleic acids research 51: D678–D689.

[15] Wang Z, Gerstein M, Snyder M (2009) Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics 10: 57–63.

[16] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. (2012) Ncbi geo: archive for functional genomics data sets—update. Nucleic acids research 41: D991–D995.

[17] Park PJ (2009) Chip–seq: advantages and challenges of a maturing technology. Nature reviews genetics 10: 669–680.

[18] Rhee HS, Pugh BF (2012) Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. Current protocols in molecular biology 100: 21–24.

[19] 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, et al. (2022) Regulondb 11.0: Comprehensive high-throughput datasets on transcriptional regulation in escherichia coli k-12. Microbial Genomics 8.

[20] Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Guyer M, Green ED (2014) The national institutes of health's big data to knowledge (bd2k) initiative: capitalizing on biomedical big data. Journal of the American Medical Informatics Association 21: 957–958.

[21] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016) The fair guiding principles for scientific data management and stewardship. Scientific data 3: 1–9.

[22] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021) Highly accurate protein structure prediction with alphafold. Nature 596: 583–589.

[23] Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, et al. (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in e. coli. PloS one 4: e7526.

[24] Pribnow D (1975) Nucleotide sequence of an rna polymerase binding site at an early t7 promoter. Proceedings of the National Academy of Sciences 72: 784–788.

[25] Mejía-Almonte C, Busby SJ, Wade JT, van Helden J, Arkin AP, Stormo GD, Eilbeck K, Palsson BO, Galagan JE, Collado-Vides J (2020) Redefining fundamental concepts of transcription initiation in bacteria. Nature Reviews Genetics 21: 699–714.

[26] Helmann JD (2019) Where to begin? sigma factors and the selectivity of transcription initiation in bacteria. Molecular microbiology 112: 335–347.

[27] Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. Nature Reviews Microbiology 2: 57–65.

[28] Zheng M, Storz G (2000) Redox sensing by prokaryotic transcription factors. Biochemical pharmacology 59: 1–6.

[29] Landis L, Xu J, Johnson RC (1999) The camp receptor protein crp can function as an osmoregulator of transcription in escherichia coli. Genes & Development 13: 3081–3091.

[30] Mukhopadhyay P, Zheng M, Bedzyk LA, LaRossa RA, Storz G (2004) Prominent roles of the norr and fur regulators in the escherichia coli transcriptional response to reactive nitrogen species. Proceedings of the National Academy of Sciences 101: 745–750.

[31] Gollnick P, Babitzke P (2002) Transcription attenuation. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1577: 240–250.

[32] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pangenome. Current opinion in genetics & development 15: 589–594.

[33] Rouli L, Merhej V, Fournier PE, Raoult D (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New microbes and new infections 7: 72–85.

[34] Wood S, Zhu K, Surujon D, Rosconi F, Ortiz-Marquez JC, van Opijnen T (2020) A pangenomic perspective on the emergence, maintenance, and predictability of antibiotic resistance. Pangenome : 169.

[35] Norsigian CJ, Fang X, Palsson BO, Monk JM (2020) Pangenome flux balance analysis toward panphenomes. The pangenome: diversity, dynamics and evolution of genomes : 219–232.

[36] Kim Y, Gu C, Kim HU, Lee SY (2020) Current status of pan-genome analysis for pathogenic bacteria. Current opinion in biotechnology 63: 54–62.

[37] 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 research 48: 10157–10163.

[38] Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J (2019) E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic acids research 47: 85–92.

[39] Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nature Reviews Genetics 7: 130–141.

[40] Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. (1997) The complete genome sequence of escherichia coli k-12. science 277: 1453–1462.

[41] 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. Nature biotechnology 27: 1043–1049.

[42] 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, et al. (2019) Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. Nucleic acids research 47: D212–D220.

[43] Thiele I, Jamshidi N, Fleming RM, Palsson BØ (2009) Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS computational biology 5: e1000312.

[44] Edwards JS, Palsson BO (2000) The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences 97: 5528–5533.

[45] 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. Nature communications 10: 5536.

[46] Hirokawa Y, Kawano H, Tanaka-Masuda K, Nakamura N, Nakagawa A, Ito M, Mori H, Oshima T, Ogasawara N (2013) Genetic manipulations restored the growth fitness of reduced-genome escherichia coli. Journal of bioscience and bioengineering 116: 52–58.

[47] Duigou S, Boccard F (2017) Long range chromosome organization in escherichia coli: The position of the replication origin defines the non-structured regions and the right and left macrodomains. PLoS genetics 13: e1006758.

[48] Bryant JA, Sellars LE, Busby SJ, Lee DJ (2014) Chromosome position effects on gene expression in escherichia coli k-12. Nucleic acids research 42: 11383–11392.

[49] Ghatak S, King ZA, Sastry A, Palsson BO (2019) The y-ome defines the 35% of escherichia coli genes that lack experimental evidence of function. Nucleic acids research 47: 2446–2454.

[50] Allen TE, Price ND, Joyce AR, Palsson BØ (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. PLoS computational biology 2: e2.

[51] Hawley DK, McClure WR (1983) Compilation and analysis of escherichia coli promoter dna sequences. Nucleic acids research 11: 2237–2255.

[52] Phaneuf PV, Gosting D, Palsson BO, Feist AM (2019) Aledb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic acids research 47: D1164–D1171.

[53] 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. Molecular systems biology 2: 2006–0008.

[54] 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 Research : gkad750.

[55] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable ai for trees. Nature machine intelligence 2: 56–67.

[56] Lioy VS, Cournac A, Marbouty M, Duigou S, Mozziconacci J, Espéli O, Boccard F, Koszul R (2018) Multiscale structuring of the e. coli chromosome by nucleoid-associated and condensin proteins. Cell 172: 771–783.

[57] Palsson BØ (2011) Systems biology: simulation of dynamic network states. Cambridge University Press.

[58] O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161: 971–987.

[59] Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics 15: 107–120.

[60] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

[61] Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the cog database. Nucleic acids research 43: D261–D269.

[62] Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, Monk JM, Zhang Z, Palsson BO (2018) ssbio: a python framework for structural systems biology. Bioinformatics 34: 2155–2157.

[63] Oliphant TE (2007) Python for scientific computing. Computing in science & engineering 9: 10–20.

[64] Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA (2011) Nupack: Analysis and design of nucleic acid systems. Journal of computational chemistry 32: 170–173.

[65] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in python. the Journal of machine Learning research 12: 2825–2830.

[66] Chen T, Guestrin C (2016) Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. NA .

[67] Giani AM, Gallo GR, Gianfranceschi L, Formenti G (2020) Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal 18: 9–19.

[68] Deng X, den Bakker HC, Hendriksen RS (2016) Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. Annual review of food science and technology 7: 353–374.

[69] Thomsen MCF, Ahrenfeldt J, Cisneros JLB, Jurtz V, Larsen MV, Hasman H, Aarestrup FM, Lund O (2016) A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. PloS one 11: e0157718.

[70] Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Current opinion in microbiology 11: 472–477.

[71] Mulligan ME, Hawley DK, Entriken R, McClure WR (1984) Escherichia coli promoter sequences predict in vitro rna polymerase selectivity. Nucleic Acids Research 12: 789–800.

[72] Collado-Vides J, Magasanik B, Gralla JD (1991) Control site location and transcriptional regulation in escherichia coli. Microbiological reviews 55: 371–394.

[73] Chen LH, Emory SA, Bricker AL, Bouvet P, Belasco JG (1991) Structure and function of a bacterial mrna stabilizer: analysis of the 5'untranslated region of ompa mrna. Journal of bacteriology 173: 4578–4586.

[74] Yamanaka K, Mitta M, Inouye M (1999) Mutation analysis of the 5' untranslated region of the cold shock cspa mrna of escherichia coli. Journal of Bacteriology 181: 6284–6291.

[75] Catoiu EA, Phaneuf P, Monk J, Palsson BO (2023) Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks. Proceedings of the National Academy of Sciences 120: e2218835120.

[76] Phaneuf PV, Jarczynska ZD, Kandasamy V, Chauhan SM, Feist AM, Palsson BO (2023) Using the e. coli alleleome in strain design. bioRxiv : 2023–09.

[77] Harke AS, Josephs-Spaulding J, Mohite OS, Chauhan SM, Ardalani O, Palsson B, Phaneuf PV (2023) Genomic insights into lactobacillaceae: Analyzing the alleleome of core pangenomes for enhanced understanding of strain diversity and revealing phylogroup-specific unique variants. bioRxiv : 2023–09.

[78] Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press.

[79] Drake JW (1991) A constant rate of spontaneous mutation in dna-based microbes. Proceedings of the National Academy of Sciences 88: 7160–7164.

[80] Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with escherichia coli. G3: Genes— Genomes— Genetics 1: 183–186.

[81] Nichols BP, Yanofsky C (1979) Nucleotide sequences of trpa of salmonella typhimurium and escherichia coli: an evolutionary comparison. Proceedings of the National Academy of Sciences 76: 5244–5248.

[82] Adelberg EA, Burns SN (1960) Genetic variation in the sex factor of escherichia coli. Journal of Bacteriology 79: 321–330.

[83] Harshman L, Riley M (1980) Conservation and variation of nucleotide sequences in escherichia coli strains isolated from nature. Journal of bacteriology 144: 560–568.

[84] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) Blast+: architecture and applications. BMC bioinformatics 10: 1–9.

[85] Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5: 1–19.

[86] Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O (2018) Clermontyping: an easy-to-use and accurate in silico method for escherichia genus strain phylotyping. Microbial genomics 4.

[87] Yoshua SB, Watson GD, Howard JA, Velasco-Berrelleza V, Leake MC, Noy A (2021) Integration host factor bends and bridges dna in a multiplicity of binding modes with varying specificity. Nucleic Acids Research 49: 8684–8698.

[88] Hyun JC, Monk JM, Palsson BO (2022) Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. BMC genomics 23: 1–18.

[89] Fu L, Niu B, Zhu Z, Wu S, Li W (2012) Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150–3152.

[90] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. (2020) Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods 17: 261–272.

[91] Ziemann M, Kaspi A, El-Osta A (2019) Digital expression explorer 2: a repository of uniformly processed rna sequencing data. Gigascience 8: giz022.

[92] Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT (2018) Flyatlas 2: a new version of the drosophila melanogaster expression atlas with rna-seq, mirna-seq and sex-specific data. Nucleic acids research 46: D809–D815.

[93] Consortium EP, et al. (2012) An integrated encyclopedia of dna elements in the human genome. Nature 489: 57.

[94] Consortium G, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, et al. (2015) The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science 348: 648–660.

[95] Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, Verendel V, Nielsen J, Töpel M, Zelezniak A (2020) Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nature Communications 11: 6141.

[96] Zhang Z, Pan Z, Ying Y, Xie Z, Adhikari S, Phillips J, Carstens RP, Black DL, Wu Y, Xing Y (2019) Deep-learning augmented rna-seq analysis of transcript splicing. Nature methods 16: 307–310.

[97] Kwon MS, Lee BT, Lee SY, Kim HU (2020) Modeling regulatory networks using machine learning for systems metabolic engineering. Current opinion in biotechnology 65: 163–170.

[98] Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research 28: 739–750.

[99] Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR (2021) Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods 18: 1196–1203.

[100] Zhang Y, Parmigiani G, Johnson WE (2020) Combat-seq: batch effect adjustment for rna-seq count data. NAR genomics and bioinformatics 2: lqaa078.

[101] Liu Q, Markatou M (2016) Evaluation of methods in removing batch effects on rna-seq data. Infect Dis Transl Med 2: 3–9.

[102] Comon P (1994) Independent component analysis, a new concept? Signal processing 36: 287–314.

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

[104] Rychel K, Sastry AV, Palsson BO (2020) Machine learning uncovers independently regulated modules in the bacillus subtilis transcriptome. Nature communications 11: 6338.

[105] Poudel S, Tsunemoto H, Seif Y, Sastry AV, Szubin R, Xu S, Machado H, Olson CA, Anand A, Pogliano J, et al. (2020) Revealing 29 sets of independently modulated genes in staphylococcus aureus, their regulators, and role in key physiological response. Proceedings of the National Academy of Sciences 117: 17228–17239.

[106] Rajput A, Tsunemoto H, Sastry AV, Szubin R, Rychel K, Sugie J, Pogliano J, Palsson BO (2022) Machine learning from pseudomonas aeruginosa transcriptomes identifies independently modulated sets of genes associated with known transcriptional regulators. Nucleic Acids Research 50: 3658–3672.

[107] Chauhan SM, Poudel S, Rychel K, Lamoureux C, Yoo R, Al Bulushi T, Yuan Y, Palsson BO, Sastry AV (2021) Machine learning uncovers a data-driven transcriptional regulatory network for the crenarchaeal thermoacidophile sulfolobus acidocaldarius. Frontiers in Microbiology 12: 753521.

[108] Yoo R, Rychel K, Poudel S, Al-Bulushi T, Yuan Y, Chauhan S, Lamoureux C, Palsson BO, Sastry A (2022) Machine learning of all mycobacterium tuberculosis h37rv rna-seq data reveals a structured interplay between metabolism, stress response, and infection. Msphere 7: e00033–22.

[109] 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 research 49: D112–D120.

[110] Du B, Olson CA, Sastry AV, Fang X, Phaneuf PV, Chen K, Wu M, Szubin R, Xu S, Gao Y, et al. (2020) Adaptive laboratory evolution of escherichia coli under acid stress. Microbiology 166: 141.

[111] Chen K, Anand A, Olson C, Sandberg TE, Gao Y, Mih N, Palsson BO (2021) Bacterial fitness landscapes stratify based on proteome allocation associated with discrete aero-types. PLoS computational biology 17: e1008596.

[112] Anand A, Chen K, Yang L, Sastry AV, Olson CA, Poudel S, Seif Y, Hefner Y, Phaneuf PV, Xu S, et al. (2019) Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. Proceedings of the National Academy of Sciences 116: 25287–25292.

[113] Anand A, Chen K, Catoiu E, Sastry AV, Olson CA, Sandberg TE, Seif Y, Xu S, Szubin R, Yang L, et al. (2020) Oxyr is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. Molecular Biology and Evolution 37: 660–667.

[114] McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO (2018) Evolution of gene knockout strains of e. coli reveal regulatory architectures governed by metabolism. Nature communications 9: 3796.

[115] 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. Metabolic Engineering 61: 360–368.

[116] Sandberg TE, Szubin R, Phaneuf PV, Palsson BO (2020) Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. Nature ecology & evolution 4: 1402–1409.

[117] Choudhary KS, Kleinmanns JA, Decker K, Sastry AV, Gao Y, Szubin R, Seif Y, Palsson BO (2020) Elucidation of regulatory modes for five two-component systems in escherichia coli reveals novel relationships. Msystems 5: e00980–20.

[118] Sastry A, Dillon N, Poudel S, Hefner Y, Xu S, Szubin R, Feist A, Nizet V, Palsson B (2020) Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility. bioRxiv : 2020–05.

[119] 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. Nature biotechnology 34: 104–110.

[120] Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO (2018) Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature communications 9: 5252.

[121] Braun V, Rehn K (1969) Chemical characterization, spatial distribution and function of a lipoprotein (murein-lipoprotein) of the e. coli cell wall: the specific effect of trypsin on the membrane structure. European Journal of Biochemistry 10: 426–438.

[122] Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635.

[123] Fleischer R, Heermann R, Jung K, Hunke S (2007) Purification, reconstitution, and characterization of the cpxrap envelope stress system of escherichia coli. Journal of Biological Chemistry 282: 8583–8593.

[124] Tschauner K, Hörnschemeyer P, Müller VS, Hunke S (2014) Dynamic interaction between the cpxa sensor kinase and the periplasmic accessory protein cpxp mediates signal recognition in e. coli. PloS one 9: e107383.

[125] Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, Feist AM, Gonzalez DJ, Palsson BO (2020) Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proceedings of the National Academy of Sciences 117: 23182–23190.

[126] Qiu S, Lamoureux C, Akbari A, Palsson BO, Zielinski DC (2022) Quantitative sequence basis for the e. coli transcriptional regulatory network. bioRxiv : 2022–02.

[127] Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, Sastry AV, Fang X, Mih N, Yang L, et al. (2018) Systematic discovery of uncharacterized transcription factors in escherichia coli k-12 mg1655. Nucleic acids research 46: 10682–10696.

[128] 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 research 49: 9696–9710.

[129] Kim GB, Gao Y, Palsson BO, Lee SY (2021) Deeptfactor: A deep learning-based tool for the prediction of transcription factors. Proceedings of the National Academy of Sciences 118: e2021171118.

[130] Rodionova IA, Gao Y, Sastry A, Yoo R, Rodionov DA, Saier Jr MH, Palsson BØ (2020) Synthesis of the novel transporter ydhc, is regulated by the ydhb transcription factor controlling adenosine and adenine uptake. bioRxiv : 2020–05.

[131] Rodionova IA, Gao Y, Monk J, Hefner Y, Wong N, Szubin R, Lim HG, Rodionov DA, Zhang Z, Saier Jr MH, et al. (2022) A systems approach discovers the role and characteristics of seven lysr type transcription factors in escherichia coli. Scientific Reports 12: 7274.

[132] Sastry AV, Hu A, Heckmann D, Poudel S, Kavvas E, Palsson BO (2021) Independent component analysis recovers consistent regulatory signals from disparate datasets. PLoS computational biology 17: e1008647.

[133] Reitzer L, Schneider BL (2001) Metabolic context and possible physiological themes of ς54-dependent genes in escherichia coli. Microbiology and Molecular Biology Reviews 65: 422–444.

[134] DeLisa MP, Wu CF, Wang L, Valdes JJ, Bentley WE (2001) Dna microarray-based identification of genes controlled by autoinducer 2-stimulated quorum sensing in escherichia coli. Journal of bacteriology 183: 5239–5247.

[135] Mehta P, Casjens S, Krishnaswamy S (2004) Analysis of the lambdoid prophage element e14 in the e. coli k-12 genome. BMC microbiology 4: 1–13.

[136] Touati D, Jacques M, Tardat B, Bouchard L, Despied S (1995) Lethal oxidative damage and mutagenesis are generated by iron in delta fur mutants of escherichia coli: protective role of superoxide dismutase. Journal of bacteriology 177: 2305–2314.

[137] Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM, Ebright RH (2004) Catabolite activator protein: Dna binding and transcription activation. Current opinion in structural biology 14: 10–20.

[138] Busby S, Ebright RH (1999) Transcription activation by catabolite activator protein (cap). Journal of molecular biology 293: 199–213.

[139] 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.

[140] Kodama Y, Shumway M, Leinonen R (2012) The sequence read archive: explosive growth of sequencing data. Nucleic acids research 40: D54–D56.

[141] Potts AH, Vakulskas CA, Pannuri A, Yakhnin H, Babitzke P, Romeo T (2017) Global role of the bacterial post-transcriptional regulator csra revealed by integrated transcriptomics. Nature communications 8: 1596.

[142] Bui TT, Selvarajoo K (2020) Attractor concepts to evaluate the transcriptome-wide dynamics guiding anaerobic to aerobic state transition in escherichia coli. Scientific Reports 10: 5878.

[143] Moore LJ, Kiley PJ (2001) Characterization of the dimerization domain in the fnr transcription factor. Journal of Biological Chemistry 276: 45744–45750.

[144] Khoroshilova N, Popescu C, Münck E, Beinert H, Kiley PJ (1997) Iron-sulfur cluster disassembly in the fnr protein of escherichia coli by o2:[4fe-4s] to [2fe-2s] conversion with loss of biological activity. Proceedings of the National Academy of Sciences 94: 6087–6092.

[145] Sutton VR, Mettert EL, Beinert H, Kiley PJ (2004) Kinetic analysis of the oxidative conversion of the [4fe-4s] 2+ cluster of fnr to a [2fe-2s] 2+ cluster. Journal of Bacteriology 186: 8018–8025.

[146] Jervis AJ, Crack JC, White G, Artymiuk PJ, Cheesman MR, Thomson AJ, Le Brun NE, Green J (2009) The o2 sensitivity of the transcription factor fnr is controlled by ser24 modulating the kinetics of [4fe-4s] to [2fe-2s] conversion. Proceedings of the National Academy of Sciences 106: 4659–4664.

[147] Salmon K, Hung Sp, Mekjian K, Baldi P, Hatfield GW, Gunsalus RP (2003) Global gene expression profiling in escherichia coli k12: the effects of oxygen availability and fnr. Journal of Biological Chemistry 278: 29837–29855.

[148] Bekker M, Alexeeva S, Laan W, Sawers G, Teixeira de Mattos J, Hellingwerf K (2010) The arcba two-component system of escherichia coli is regulated by the redox state of both the ubiquinone and the menaquinone pool. Journal of bacteriology 192: 746–754.

[149] Van Beilen JW, Hellingwerf KJ (2016) All three endogenous quinone species of escherichia coli are involved in controlling the activity of the aerobic/anaerobic response regulator arca. Frontiers in microbiology 7: 1339.

[150] Iuchi S, Lin E (1988) arca (dye), a global regulatory gene in escherichia coli mediating repression of enzymes in aerobic pathways. Proceedings of the National Academy of Sciences 85: 1888–1892.

[151] Iuchi S, Lin E (1991) Adaptation of escherichia coli to respiratory conditions: regulation of gene expression. Cell 66: 5–7.

[152] Gunsalus R, Park SJ (1994) Aerobic-anaerobic gene regulation in escherichia coli: control by the arcab and fnr regulons. Research in microbiology 145: 437–450.

[153] Mills SA, Marletta MA (2005) Metal binding characteristics and role of iron oxidation in the ferric uptake regulator from escherichia coli. Biochemistry 44: 13553–13559.

[154] Beauchene NA, Myers KS, Chung D, Park DM, Weisnicht AM, Keleş S, Kiley PJ (2015) Impact of anaerobiosis on expression of the iron-responsive fur and ryhb regulons. MBio 6: 10–1128.

[155] Nunoshiba T, Hidalgo E, Amabile Cuevas C, Demple B (1992) Two-stage control of an oxidative stress regulon: the escherichia coli soxr protein triggers redox-inducible expression of the soxs regulatory gene. Journal of bacteriology 174: 6054–6060.

[156] Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G (2001) Dna microarray-mediated transcriptional profiling of the escherichia coli response to hydrogen peroxide. Journal of bacteriology 183: 4562–4570.

[157] Stephenson M, Stickland LH (1932) Hydrogenlyases: Bacterial enzymes liberating molecular hydrogen. Biochemical Journal 26: 712.

[158] Lim HG, Rychel K, Sastry AV, Bentley GJ, Mueller J, Schindel HS, Larsen PE, Laible PD, Guss AM, Niu W, et al. (2022) Machine-learning from pseudomonas putida kt2440 transcriptomes reveals its transcriptional regulatory network. Metabolic Engineering 72: 297–310.

[159] Rodionova IA, Gao Y, Sastry A, Hefner Y, Lim HG, Rodionov DA, Saier Jr MH, Palsson BO (2021) Identification of a transcription factor, punr, that regulates the purine and purine nucleoside transporter punc in e. coli. Communications biology 4: 991.

[160] 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.

[161] Anand A, Patel A, Chen K, Olson CA, Phaneuf PV, Lamoureux C, Hefner Y, Szubin R, Feist AM, Palsson BO (2022) Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity. Nature Communications 13: 3682.

[162] Kavvas ES, Long CP, Sastry A, Poudel S, Antoniewicz MR, Ding Y, Mohamed ET, Szubin R, Monk JM, Feist AM, et al. (2022) Experimental evolution reveals unifying systems-level adaptations but diversity in driving genotypes. Msystems 7: e00165–22.

[163] Sastry AV, Dillon N, Anand A, Poudel S, Hefner Y, Xu S, Szubin R, Feist AM, Nizet V, Palsson B (2021) Machine learning of bacterial transcriptomes reveals responses underlying differential antibiotic susceptibility. Msphere 6: e00443–21.

[164] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. Nature biotechnology 35: 316–319.

[165] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome biology 10: 1–10.

[166] Wang L, Wang S, Li W (2012) Rseqc: quality control of rna-seq experiments. Bioinformatics 28: 2184–2185.

[167] Liao Y, Smyth GK, Shi W (2014) featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923–930.

[168] Ewels P, Magnusson M, Lundin S, Käller M (2016) Multiqc: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

[169] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology 15: 1–21.

[170] Hyvarinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE transactions on Neural Networks 10: 626–634.

[171] McConn JL, Lamoureux CR, Poudel S, Palsson BO, Sastry AV (2021) Optimal dimensionality selection for independent component analysis of transcriptomic data. BMC bioinformatics 22: 1–13.