

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Decomposing the symphony of Escherichia coli gene expression

Permalink

<https://escholarship.org/uc/item/6nz923zg>

Author

Sastry, Anand Varun

Publication Date

2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Decomposing the symphony of *Escherichia coli* gene expression

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

in

Bioengineering

by

Anand Varun Sastry

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Rob Knight
Professor Andrew McCulloch
Professor Joseph Pogliano
Professor Kun Zhang

2020

Copyright
Anand Varun Sastry, 2020
All rights reserved.

The dissertation of Anand Varun Sastry is approved, and
it is acceptable in quality and form for publication on
microfilm and electronically:

Chair

University of California San Diego

2020

DEDICATION

To my family, for their love and support throughout my life

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
Acknowledgements	xi
Vita	xiii
Abstract of the Dissertation	xv
Chapter 1	
Introduction	1
1.1 Transcriptional regulation in <i>E. coli</i>	2
1.2 Big Biological Data and Machine Learning	3
1.3 Independent Component Analysis	5
1.4 Thesis Outline	7
1.5 References	7
Chapter 2	
The <i>E. coli</i> transcriptome mostly consists of independently modulated groups of genes	11
2.1 Background	12
2.2 Results	14
2.2.1 ICA extracts regulatory signals from expression data	14
2.2.2 Validation of i-modulon-regulator relationships	23
2.2.3 ICA reveals independent modules within the PurR regulon	26
2.2.4 ICA provides answers to unasked questions	29
2.2.5 Two i-modulons characterize the ‘Fear-Greed’ trade-off	33
2.2.6 An i-modulon discriminates between two <i>E. coli</i> strains	35
2.2.7 I-modulons explain expression changes across <i>E. coli</i> strains	37
2.3 Discussion	38
2.4 Methods	39
2.4.1 RNA extraction and library preparation	39
2.4.2 ChIP-exo preparation	40
2.4.3 ChIP-exo Processing	41
2.4.4 Compilation of PRECISE	42
2.4.5 Compilation of the reported <i>E. coli</i> regulatory network	42
2.4.6 Computing robust independent components	43
2.4.7 Determination of the gene coefficient threshold	44
2.4.8 Associating regulators to i-modulons	46
2.4.9 Cumulative explained variance for ICA	46

	2.4.10 Comparison of microarray data and PRECISE	46
	2.4.11 Differential activity analysis	48
	2.4.12 I-modulon summation	48
	2.4.13 Multiple strain RNA-seq processing and projection	49
	2.5 References	51
Chapter 3	Matrix factorization recovers consistent regulatory signals from disparate datasets	59
	3.1 Background	60
	3.2 Results	62
	3.2.1 ICA extracts i-modulons from five transcriptomic datasets . . .	62
	3.2.2 I-modulon structure is conserved across transcriptomic datasets	66
	3.2.3 ICA finds similar structures in the proteome	69
	3.2.4 Data integration reveals increased resolution and novel insights	73
	3.3 Discussion	77
	3.4 Methods	79
	3.4.1 RNA-seq and microarray data processing	79
	3.4.2 Proteomics data processing	81
	3.4.3 Independent Component Analysis	81
	3.4.4 Characterizing i-modulons	82
	3.4.5 Comparing i-modulon structures	83
	3.4.6 Linear regression of i-modulons	83
	3.5 References	84
Chapter 4	Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility	90
	4.1 Background	91
	4.2 Results	92
	4.2.1 Antibiotic efficacy varies across media compositions	92
	4.2.2 A few mechanisms can explain large transcriptomic perturbations	95
	4.2.3 Transcriptional markers of ‘greedy’ growth in rich media	97
	4.2.4 Subinhibitory antibiotic treatment decelerates respiration . . .	100
	4.2.5 Ciprofloxacin has diverging effects on Fur i-modulons	102
	4.3 Discussion	104
	4.4 Methods	105
	4.4.1 Bacterial strains and growth conditions	105
	4.4.2 Antibiotic susceptibility testing	105
	4.4.3 Bactericidal Assays	106
	4.4.4 RNA-seq expression profiling and processing	106
	4.4.5 Independent component analysis	106
	4.5 References	107
Chapter 5	Conclusions	111

Appendix A	The <i>E. coli</i> transcriptome mostly consists of independently modulated groups of genes - Supplementary Information	114
A.1	Supplementary Figures	115
A.2	References	122
Appendix B	Descriptive characteristics of the 92 i-modulons extracted from PRECISE .	123
B.1	Automated characterization of i-modulons	123
B.1.1	Motif Search	125
B.2	References	127
Appendix C	Matrix factorization recovers consistent regulatory signals from disparate datasets	220
C.1	Supplementary Figures	221
Appendix D	Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility	227
D.1	Supplementary Figures	228

LIST OF FIGURES

Figure 1.1:	Simplified illustration of the TRN	2
Figure 1.2:	Current state of <i>E. coli</i> TRN and transcriptomic data	3
Figure 1.3:	Matrix decomposition of gene expression datasets	5
Figure 1.4:	ICA extracts individual signals from mixed recordings	6
Figure 2.1:	Independent component analysis and blind source separation	16
Figure 2.2:	Characterization of independent components	17
Figure 2.3:	I-modulons are enriched with metabolic pathways	22
Figure 2.4:	Validation of i-modulon-regulator relationships	24
Figure 2.5:	Validation of i-modulon activities	26
Figure 2.6:	ICA reveals independent modules within the PurR regulon	27
Figure 2.7:	ICA provides answers to unasked questions	31
Figure 2.8:	Two i-modulons characterize the ‘Fear vs. Greed’ trade-off	34
Figure 2.9:	I-modulons explain expression changes across <i>E. coli</i> strains	36
Figure 3.1:	Schematic illustration of the workflow applied to each data set	63
Figure 3.2:	Characterization of i-modulons derived from five independent gene expression datasets	65
Figure 3.3:	The i-modulon structure is conserved across five datasets.	67
Figure 3.4:	Comparison of IC gene coefficients for i-modulons enriched with genes in the CysB regulon	70
Figure 3.5:	ICA elucidates similar structures in the <i>E. coli</i> transcriptome and proteome	72
Figure 3.6:	Decomposition of combined compendium results in increased resolution	74
Figure 4.1:	Inoculum cell density affects antibiotic efficacy	94
Figure 4.2:	Global comparative analysis of the <i>E. coli</i> transcriptome across three media compositions	96
Figure 4.3:	Mechanisms underlying the complex transcriptional response to different media	98
Figure 4.4:	Consistent i-modulon responses to antibiotic treatment	101
Figure 4.5:	I-modulon responses to ciprofloxacin treatment	103
Figure 5.1:	ICA presents a white-box model of the TRN	112
Figure A.1:	Overview of the PRECISE Dataset	115
Figure A.2:	Overview of i-modulons	116
Figure A.3:	Validation of i-modulon-regulator relationships	117
Figure A.4:	Additional properties of i-modulons	118
Figure A.5:	Correlations related to the fear-greed trade-off	119
Figure A.6:	Significance thresholds for components and differential activity	119
Figure A.7:	Annotated example of an i-modulon dashboard (Part 1)	120
Figure A.8:	Annotated example of an i-modulon dashboard (Part 2)	121
Figure C.1:	RBH graph containing all edges from the five datasets	221
Figure C.2:	The Genomic i-modulon found in more than one dataset	222
Figure C.3:	Investigation of an irregular i-modulon cluster	223

Figure C.4: ICA finds similar structures in the transcriptome and proteome	224
Figure C.5: ICA finds similar structures in the transcriptome and proteome	225
Figure C.6: Effects of data integration on i-modulon structure and activity	226
Figure D.1: Bar charts for additional i-modulon activities	228
Figure D.2: Heatmap of i-modulon activities after antibiotic treatment	228
Figure D.3: Bar chart of fear-greed i-modulon activities	229
Figure D.4: Bar chart of nutrient-related i-modulon activities	229
Figure D.5: Venn diagram of genes in the Fur i-modulons	230

LIST OF TABLES

Table 2.1: Overview of i-modulons extracted from PRECISE	18
Table 3.1: Summary of transcriptomic datasets	63
Table 4.1: MIC of nine antibiotics on three media. Antibiotics in bold were selected for further analysis.	93

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Professor Bernhard Palsson for his mentorship and guidance over the past five and a half years. His scientific insight and practical advice provided invaluable motivation and helped me understand the value of my work. Despite his busy schedule he has always provided support when needed, and has been very in tune with the challenges we face in grad school. I cannot thank him enough for the opportunities I've been given.

I have been fortunate to have had many inspiring mentors at the SBRG. First, I would like to thank Liz Brunk and Dan Zielinski for welcoming me into the lab, and introducing me to machine learning and the systems perspective of biology. Zak King and Jon Monk were my first contact with my lab, and they have provided critical scientific and career advice throughout this process. I also thank Laurence Yang and David Heckmann for being patient and thoughtful role models in scientific leadership.

So much of this thesis was critically dependent on high-quality RNA-seq data that were generated in our lab. I am indebted to Richard Szubin, Ying Hefner, Julia Xu, Ye Gao, Donghyuk Kim, Sangwoo Seo, and Amitesh Anand for preparing this data. This thesis would not exist without their contributions.

I am grateful for the culture of innovation and collaboration fostered by everyone in the SBRG. There are so many colleagues and friends to thank here, but I would especially like to thank JC, Colton, Yara, Saugat, Erol, CJ, Patrick, Bin, Muyao, Justin, and Sonal for daily invigorating coffee chats and relaxing outdoor lunches. I must also thank Kevin, Jacob, Cam, and Katherine for their assistance while I prepared my thesis. In addition, I would like to thank my grad school cohort for helping me maintain a work-life balance.

I am especially thankful for the love, inspiration, and motivation from my parents and my sister throughout my life. I could not have made it this far without you, and this thesis is dedicated to you.

Finally, I would like to thank the love of my life, Rachel, for everything you do, small and large, to make me feel loved and special. I am excited for our next chapter together.

I would also like to thank my funding sources that have supported this work. These include the Novo Nordisk Foundation (NNF10CC1016517) and the NIH NIAID grant (1-U01-AI124316).

Chapter 2, in part, is a reprint of the material as appears in: **AV Sastry**, Y Gao, R Szubin, Y Hefner, S Xu, D Kim, KS Choudhary, L Yang, ZA King, and BO Palsson. 2019. “The *Escherichia coli* transcriptome mostly consists of independently regulated modules” *Nature Communications*. 10 (5536). The dissertation author was the primary author.

Chapter 3, in part, is currently being prepared for submission for publication: **AV Sastry**, A Hu, D Heckmann, S Poudel, E Kavvas, and BO Palsson. 2019. “Matrix factorization recovers consistent regulatory signals from disparate datasets”. The dissertation author is the primary author.

Chapter 4, in part, is currently being prepared for submission for publication. **AV Sastry***, N Dillon*, S Poudel, Y Hefner, S Xu, R Szubin, AM Feist, V Nizet, BO Palsson “Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility”. The dissertation author is one of the two primary authors.

VITA

2014	Bachelor of Science in Chemical Engineering, Double Major in Biomedical Engineering, Carnegie Mellon University
2020	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*. 2019;10: 5536.

Anand A, Chen K, Yang L, **Sastry AV**, Olson CA, Poudel S, et al. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. *Proceedings of the National Academy of Science*. 2019;116: 25287–25292.

Anand A, Chen K, Catoiu E, **Sastry AV**, Olson CA, Sandberg TE, et al. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. *Molecular Biology and Evolution*. 2019.

Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, Kim D, **Sastry AV** et al. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. *Proceedings of the National Academy of Science*. 2019;116: 14368–14373.

Lachance J-C, Lloyd CJ, Monk JM, Yang L, **Sastry AV**, Seif Y, et al. BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Computational Biology*. 2019;15: e1006971.

Anand A, Olson CA, Yang L, **Sastry AV**, Catoiu E, Choudhary KS, et al. Pseudogene repair driven by selection pressure applied in experimental evolution. *Nature Microbiology*. 2019;4: 386–389.

Ghatak S, King ZA, **Sastry A**, Palsson BO. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Research*. 2019;47: 2446–2454.

Fang X, Monk JM, Mih N, Du B, **Sastry AV**, Kavvas E, et al. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Systems Biology*. 2018;12: 66.

Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, **Sastry AV** et al. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. *Nucleic Acids Research*. 2018.

Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, Gatto F, Nilsson A, Gonzalez GAP, Aurich MK, Prlić A, **Sastry A** et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*. 2018;36: 272.

Mih N, Brunk E, Chen K, Catoiu E, **Sastry A**, Kavvas E, et al. ssbio: a Python framework for structural systems biology. *Bioinformatics*. 2018;34: 2155–2157.

Monk JM*, Lloyd CJ*, Brunk E, Mih N, **Sastry A**, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. *Nature Biotechnology*. 2017;35: 904–908.

Fang X*, **Sastry A***, Mih N, Kim D, Tan J, Yurkovich JT, et al. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences*. 2017.

Sastry A, Monk J, Tegel H, Uhlén M, Palsson BO, Rockberg J, et al. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. *Bioinformatics*. 2017.

Ebrahim A*, Brunk E*, Tan J*, O’Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, **Sastry A** et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications*. 2016;7: 13091.

* equal contribution

ABSTRACT OF THE DISSERTATION

Decomposing the symphony of *Escherichia coli* gene expression

by

Anand Varun Sastry

Doctor of Philosophy in Bioengineering

University of California San Diego, 2020

Professor Bernhard Ø. Palsson, Chair

Bacteria respond and adapt to dynamic environments by altering their gene expression through a complex Transcriptional Regulatory Network (TRN). Advances in sequencing technologies have accelerated the generation of large RNA sequencing datasets that can be leveraged to probe the TRN. Here, we analyze large gene expression datasets using Independent Component Analysis (ICA), an unsupervised machine learning algorithm developed to separate source signals (e.g. individual instruments) from a set of mixed signals (e.g. recordings of an orchestra). First, we compile a high-quality RNA-seq compendium containing over 250 expression profiles for the model bacteria *Escherichia coli*. We apply ICA to decompose this compendium into indepen-

dently modulated sets of genes, termed i-modulons, that represent the source signals comprising the transcriptome. Of the 92 i-modulons, 61 capture the genome-wide targets of characterized transcriptional regulators with high accuracy. ICA simultaneously estimates the activity of each transcriptional regulator in each growth condition. We show that this representation of the TRN can be used to predict and validate new regulatory interactions, characterize mutations in transcription factors, and provide a basis to compare gene expression across multiple *E. coli* strains. Next, we show that the underlying structure of the *E. coli* transcriptome, as determined by ICA, is conserved across multiple independent RNA-seq and microarray datasets. We subsequently combine the datasets into a compendium containing over 800 expression profiles, and discover that the ICA-based structure is still maintained upon data integration. Echoes of this structure were also found in two proteomics datasets, accelerating biological discovery through multi-omics analysis. Finally, we investigate how gene expression links environmental conditions to antibiotic efficacy. We show that i-modulons simplify the analysis of complex transcriptional changes to enable rapid characterization of cellular states beyond the transcriptome. As a whole, this body of work introduces ICA as a compelling tool to integrate and understand large gene expression datasets.

Chapter 1

Introduction

Bacteria survive and thrive on diverse environments by coordinating their gene expression through a complex transcriptional regulatory network (TRN). For example, the gut-colonizing bacteria *Escherichia coli* must respond to fluctuating nutrient and oxygenation levels by modulating its gene expression state, or transcriptome. The TRN is an information processing unit; it senses the cell's environment and intracellular state, integrates this information, and produces a transcriptome intended to optimize the cell's chance for survival (Figure 1.1). Recent advances in high-throughput sequencing technologies have enabled rapid generation of high-quality transcriptomes that can be leveraged to reverse engineer TRNs.

Understanding microbial TRNs is crucial for many issues in bioengineering and healthcare. Transcription factors have been implicated in controlling virulence in pathogenic organisms [1, 2], evolution of antibiotic resistance [3, 4], and engineering of microbial strains for bio-production [5, 6]. A complete understanding of microbial gene expression and regulation would allow us to explain and predict the cell's transcriptomic responses to these complex environmental states.

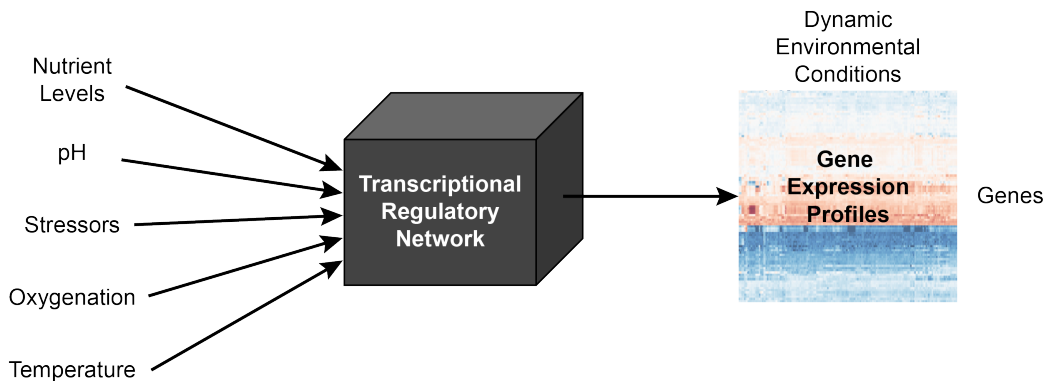


Figure 1.1: Simplified illustration of the TRN

1.1 Transcriptional regulation in *E. coli*

Microbial TRNs contain a wide range of transcriptional regulators, including proteins (transcription factors and sigma factors [7]) and RNA elements (small RNAs [8], riboswitches [9], and transcriptional attenuators [10]). High-throughput methods such as chromatin immunoprecipitation (ChIP) [11], can be used to determine the specific binding sites of transcription factors or sigma factors on a genome-scale. However, time-intensive biochemical assays are necessary to identify the molecular signals sensed by these regulators or to interrogate RNA-level regulators.

The model organism *Escherichia coli* is one of the best characterized organisms, with known functions for over 65% of the genes in the laboratory strain MG1655 [12]. Thanks to the dedicated work of past scientists, there are now over 7,000 known interactions between transcriptional regulators and genes in *E. coli*, which have been compiled into RegulonDB [13]. A brief examination of this database shows that over 75% of genes in RegulonDB are known targets of two or fewer TFs (Figure 1.2a). However, multiple studies have shown that mere binding information is still not enough to predict or even explain global changes in gene expression

[14, 15], highlighting the need for novel methods and models to resolve microbial gene regulation.

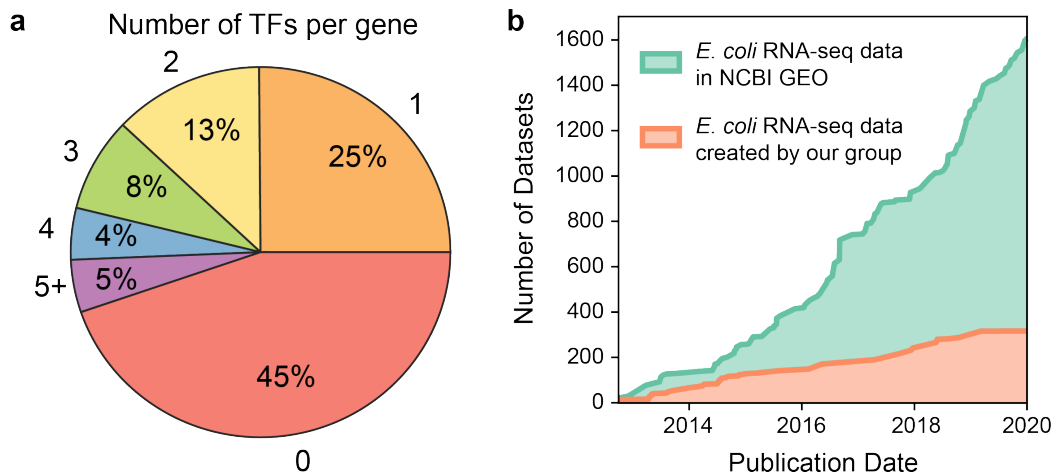


Figure 1.2: Current state of *E. coli* TRN and transcriptomic data. (a) Distribution of TFs per gene in *E. coli*. (b) Growth of *E. coli* transcriptomic data. Only datasets for *E. coli* MG1655 and BW25113 are shown

1.2 Big Biological Data and Machine Learning

A promising route to exploring complex biological systems is through statistical and machine learning analysis of big biological data. In 2014, the NIH started the “Big Data to Knowledge” initiative to promote the development of analytical methods and organizational tools for making sense of the recent explosion of data [16]. The NCBI Gene Expression Omnibus (GEO) [17] exemplifies this data deluge, as it has increased from 2 million to 3 million transcriptomes in the year 2019 [18]. This exponential growth is also reflected in *E. coli*-specific datasets, which is in part driven by our research group (Figure 1.2b). Machine learning methods can be applied to these datasets to extract general rules and patterns from data.

Machine learning can be broadly categorized into supervised learning and unsupervised learning. Supervised learning is predictive; given a set of inputs and outputs, a supervised

learning algorithm can learn rules to predict the output based on new inputs. The simplest example of supervised learning is linear regression. My first publication, which was excluded from this thesis, involved training multiple machine learning algorithms to predict the expression of short peptides from multi-dimensional properties of their amino acid and nucleotide sequences [19].

Sometimes, there are no expected outcomes from a dataset. In these cases, unsupervised learning can find structures or patterns that shed light on processes represented by the data. For example, clustering of gene expression data can identify co-expressed genes that share similar functions. In my second publication, my co-authors and I used unsupervised learning to investigate the completeness and robustness of *E. coli*'s TRN [14]. In the process, we found that systemic biases could affect transcriptomic datasets compiled from multiple data sources, even after correcting for batch effects. This finding underscored the need for a large high-quality, self-consistent gene expression compendium.

Multiple unsupervised learning algorithms fall under the general category of matrix decomposition, including principal component analysis, independent component analysis, and non-negative matrix factorization. These methods decompose a high-dimensional matrix into two lower-dimensional matrices (Figure 1.3). When applied to gene expression data, one matrix contains relationships between genes, and the other contains sample-specific weights or activities [20]. This thesis specifically focuses on the application of independent component analysis (ICA) to microbial gene expression datasets.

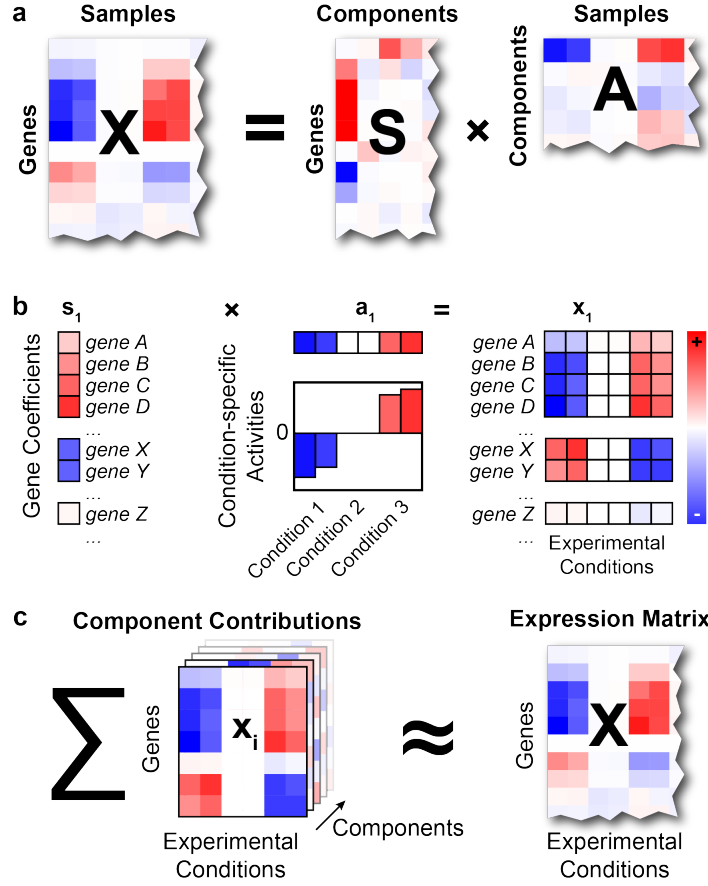


Figure 1.3: Matrix decomposition of gene expression datasets. (a) General equation of matrix decomposition. Variable names are specific to ICA. (b) Each component (column of S) contains a coefficient for each gene. These coefficients are scaled by the component's condition-specific activities (row in A) to form the component's contribution to the transcriptomic compendium (b) The sum of the contributions from the all components reconstructs the original matrix.

1.3 Independent Component Analysis

ICA was developed to solve the problem of blind source separation [21]. This can be allegorized using an orchestra warming up, where each instrument is playing its own melody (Figure 1.4). If we place many microphones throughout the stage, each microphone will record a different linear combination of instruments depending on their relative locations. Using only this set of mixed signals, and no additional information, ICA can isolate the sound of each individual

instrument and decompose the orchestra into its individual voices. In addition, ICA recovers the relative volumes of these individual signals, reflecting the distances of each microphone from each instrument.

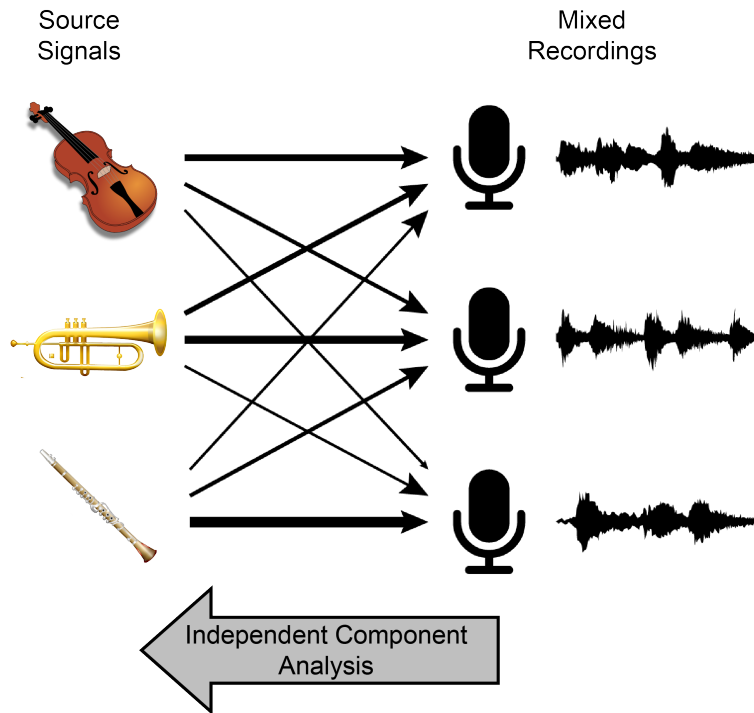


Figure 1.4: ICA extracts individual voices from mixed signals.

Mathematically, ICA is represented by the equation $\mathbf{X} = \mathbf{SA}$, where $\mathbf{X}$ is the matrix of mixed measurements, $\mathbf{S}$ is the matrix of individual signals, and $\mathbf{A}$ is the matrix containing the relative activities or contributions of each individual signal to each mixed measurement (Figure 1.3). ICA relies on the assumption that individual signals are statistically independent from each other, such that each signal has no effect on or relationship to the other signals. This problem can be further simplified, since the statistical property of non-gaussianity is enough to guarantee independence [22]. Through these properties a fast algorithm for ICA, aptly named FastICA,

was developed [23].

ICA has been shown to recover biologically-relevant signals from gene expression data [24–30]. However, these prior studies could not conclusively state the biological source underlying most of their independent components.

1.4 Thesis Outline

In this thesis, we present ICA as a tool to extract TRNs from large transcriptomic compendia. Each chapter in this thesis represents one of the matrices in the fundamental ICA equation. In the next chapter, we introduce the Precision RNA-seq Expression Compendium for Independent Signal Extraction, or PRECISE, which is represented by the $\mathbf{X}$ matrix. This dataset is used to demonstrate a variety of properties and insights revealed by application of ICA. In the third chapter, we investigate the robustness of the $\mathbf{S}$ matrix and show that similar structures are found across disparate datasets. Finally, the fourth chapter expands upon the utility of the $\mathbf{A}$ matrix, demonstrating that ICA of transcriptomics data can provide cellular characterization beyond the transcriptome. Taken together, the results presented in this thesis introduce ICA as an integral method for understanding microbial gene expression.

1.5 References

1. Raghavan, S., Manzanillo, P., Chan, K., Dovey, C. & Cox, J. S. Secreted transcription factor controls *Mycobacterium tuberculosis* virulence. en. *Nature* **454**, 717–721 (Aug. 2008).
2. Kachroo, P., Eraso, J. M., Beres, S. B., Olsen, R. J., Zhu, L., Nasser, W., Bernard, P. E., Cantu, C. C., Saavedra, M. O., Arredondo, M. J., Strobe, B., Do, H., Kumaraswami, M., Vuopio, J., Gröndahl-Yli-Hannuksela, K., Kristinsson, K. G., Gottfredsson, M., Pesonen, M., Pensar, J., Davenport, E. R., Clark, A. G., Corander, J., Caugant, D. A., Gaini, S., Magnussen, M. D., Kubiak, S. L., Nguyen, H. A. T., Long, S. W., Porter, A. R., DeLeo, F. R. & Musser, J. M. Integrated analysis of population genomics, transcriptomics and

- virulence provides novel insights into *Streptococcus pyogenes* pathogenesis. en. *Nat. Genet.* **51**, 548–559 (Mar. 2019).
3. Lázár, V., Martins, A., Spohn, R., Daruka, L., Grézal, G., Fekete, G., Számel, M., Jangir, P. K., Kintsés, B., Csörgő, B., Nyerges, Á., Györkei, Á., Kincses, A., Dér, A., Walter, F. R., Deli, M. A., Urbán, E., Hegedűs, Z., Olajos, G., Méhi, O., Bálint, B., Nagy, I., Martinek, T. A., Papp, B. & Pál, C. Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. en. *Nat Microbiol* **3**, 718–731 (June 2018).
 4. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. en. *Nat. Rev. Microbiol.* **13**, 42–51 (Jan. 2015).
 5. Mohamed, E. T., Wang, S., Lennen, R. M., Herrgård, M. J., Simmons, B. A., Singer, S. W. & Feist, A. M. Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. en. *Microb. Cell Fact.* **16**, 204 (Nov. 2017).
 6. Choe, D., Lee, J. H., Yoo, M., Hwang, S., Sung, B. H., Cho, S., Palsson, B., Kim, S. C. & Cho, B.-K. Adaptive laboratory evolution of a genome-reduced *Escherichia coli*. en. *Nat. Commun.* **10**, 935 (Feb. 2019).
 7. Ishihama, A. Functional Modulation of *Escherichia Coli* RNA Polymerase. *Annu. Rev. Microbiol.* **54**, 499–518 (Oct. 2000).
 8. Gottesman, S. & Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. en. *Cold Spring Harb. Perspect. Biol.* **3** (Dec. 2011).
 9. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. en. *Trends Biochem. Sci.* **29**, 11–17 (Jan. 2004).
 10. Yanofsky, C. Transcription attenuation. en. *J. Biol. Chem.* **263**, 609–612 (Jan. 1988).
 11. Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* **100**, 21–24 (2012).
 12. Ghatak, S., King, Z. A., Sastry, A. & Palsson, B. O. The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. en. *Nucleic Acids Res.* **47**, 2446–2454 (Mar. 2019).
 13. Santos-Zavaleta, A., Sánchez-Pérez, M., Salgado, H., Velázquez-Ramírez, D. A., Gama-Castro, S., Tierrafría, V. H., Busby, S. J. W., Aquino, P., Fang, X., Palsson, B. O., Galagan, J. E. & Collado-Vides, J. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. en. *BMC Biol.* **16**, 91 (Aug. 2018).
 14. Fang, X., Sastry, A., Mih, N., Kim, D., Tan, J., Yurkovich, J. T., Lloyd, C. J., Gao, Y., Yang, L. & Palsson, B. O. Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. en. *Proc. Natl. Acad. Sci. U. S. A.* (Sept. 2017).

15. Larsen, S. J., Röttger, R., Schmidt, H. H. H. W. & Baumbach, J. E. coli gene regulatory networks are inconsistent with gene expression data. en. *Nucleic Acids Res.* **47**, 85–92 (Jan. 2019).
16. *Big Data to Knowledge — NIH Common Fund* <https://commonfund.nih.gov/bd2k>. Accessed: 2020-1-31.
17. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (Jan. 2013).
18. geo. *GEO Summary - GEO - NCBI* <https://www.ncbi.nlm.nih.gov/geo/summary>. Accessed: 2020-1-31.
19. Sastry, A., Monk, J., Tegel, H., Uhlén, M., Palsson, B. O., Rockberg, J. & Brunk, E. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. en. *Bioinformatics* (Apr. 2017).
20. Stein-O’Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y. & Fertig, E. J. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (Oct. 2018).
21. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36**, 287–314 (Apr. 1994).
22. Hyvärinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* en (John Wiley & Sons, Apr. 2004).
23. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. en. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999).
24. Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T. & Huang, X. A review of independent component analysis application to microarray gene expression data. en. *Biotechniques* **45**, 501–520 (Nov. 2008).
25. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. en. *Bioinformatics* **18**, 51–60 (Jan. 2002).
26. Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R. & Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. en. *PLoS Comput. Biol.* **3**, e161 (Aug. 2007).
27. Martoglio, A.-M., Miskin, J. W., Smith, S. K. & MacKay, D. J. C. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. en. *Bioinformatics* **18**, 1617–1624 (Dec. 2002).
28. Frigyesi, A., Veerla, S., Lindgren, D. & Höglund, M. Independent component analysis reveals new and biologically significant structures in micro array data. en. *BMC Bioinformatics* **7**, 290 (June 2006).

29. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebre, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A. & Radvanyi, F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. en. *Cell Rep.* **9**, 1235–1245 (Nov. 2014).
30. Engreitz, J. M., Daigle Jr, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. en. *J. Biomed. Inform.* **43**, 932–944 (Dec. 2010).

Chapter 2

The *E. coli* transcriptome mostly consists of independently modulated groups of genes

Underlying cellular responses is a transcriptional regulatory network (TRN) that modulates gene expression. A useful description of the TRN would decompose the transcriptome into targeted effects of individual transcriptional regulators. Here, we apply unsupervised machine learning to a diverse compendium of nearly 300 high-quality *Escherichia coli* RNA-seq datasets to identify 92 statistically independent signals that modulate the expression of specific gene sets. We show that 61 of these transcriptomic signals represent the effects of currently characterized transcriptional regulators. Condition-specific activation of signals is validated by exposure of *E. coli* to new environmental conditions. The resulting decomposition of the transcriptome provides: a mechanistic, systems-level, network-based explanation of responses to environmental

and genetic perturbations; a guide to gene and regulator function discovery; and a basis for characterizing transcriptomic differences in multiple strains. Taken together, our results show that signal summation describes the composition of a model prokaryotic transcriptome.

2.1 Background

The transcriptional regulatory network (TRN) senses and integrates complex environmental and intracellular information to coordinate gene expression of a cell. Reverse engineering the TRN informs how an organism responds to diverse stresses and unfamiliar environments [1–3]. A fully characterized TRN would enable the prediction and mechanistic explanation of an organism’s dynamic adaptation to environmental or genetic perturbations.

Reconstruction of a genome-scale TRN requires a substantial number of experiments to integrate the binding sites for each regulator and characterize their activities [4, 5]. Unlike eukaryotic TRNs, which contain highly-connected co-associations [6], prokaryotic TRNs exhibit a simpler structure; over 75% of genes in the model bacteria *Escherichia coli* are known targets of two or fewer TFs (Figure 1.1) [7].

The TRN structure is encoded in the genome as regulator binding sites and is invariant to environmental dynamics. However, environmental and genetic perturbations alter the activity states of transcriptional regulators to change their DNA binding affinity [8], which in turn modulates the transcriptome in a condition-specific manner [9]. Thus, a measured expression profile reflects a combination of the activity of all transcriptional regulators under the examined condition. This poses the fundamental deconvolution challenge of separating the condition-invariant network structure from its condition-dependent expression state on a genome scale.

Compendia of microarray expression profiles have been leveraged to infer TRNs by iden-

tifying shared patterns across gene expression profiles, rather than using direct DNA-TF binding information [10, 11]. Many inference methods define groups of genes, or modules, with similar expression profiles that are often functionally related or co-expressed. Recently, a comprehensive review of 42 module detection methods showed that independent component analysis (ICA), a signal deconvolution algorithm, outperformed all other algorithms in identifying groups of coregulated genes [12].

ICA is a blind source separation algorithm used to deconvolute mixed signals into their individual sources and determine their relative strengths [13]. Prior application of ICA to microarray expression data [14] has identified co-expressed, functionally-related gene sets [15–17] that often map to metabolic pathways [18, 19]. A major advantage of decomposition-based module detection algorithms, such as singular value decomposition (SVD) and ICA, over clustering or network inference methods is that decomposition-based methods detect gene modules while simultaneously computing the context-specific activity levels for these gene modules [12, 20]. The overall expression levels, or activities of ICA-derived gene sets have been leveraged to classify tumor samples [21–23] and connect transcriptional modules to disease states [24].

Although the aforementioned studies showed that ICA tends to identify biologically-relevant transcriptional modules, the majority of gene modules remain uncharacterized, limiting the utility of ICA-derived results to interpret biological responses. Here, we overcome this limitation by applying ICA to a high-quality RNA-seq compendium for the well-characterized model organism *E. coli*. We find that 66% of ICA-derived gene sets clearly represent the effects of transcriptional regulators, and propose biological or genetic explanations for an additional 27% of gene sets (leaving only 7% uncharacterized). Our approach relies on: (1) the availability of high-quality, self-consistent, and condition-rich RNA-seq expression profiling datasets; (2) the

use of ICA to concurrently identify regulator targets and activities; and (3) experimental validation through the association of inferred regulator targets with observed molecular interactions and through successful prediction of gene module activation. The elucidated TRN structure deconvolutes transcriptomic responses of *E. coli* into a summation of condition-specific effects of individual transcriptional regulators.

2.2 Results

2.2.1 ICA extracts regulatory signals from expression data

In order to extract regulatory interactions from expression data, diverse conditions must be profiled to discriminate between the effects of transcriptional regulators. Previous studies have compiled transcriptomics data from independent research groups to study the transcriptional states and regulation of *E. coli* [10, 25–28]. Even after resolving the significant normalization challenge with such disparate datasets, many sources of variation remain that obscure biological signals [29–31]. These datasets mostly contain microarray data; RNA sequencing (RNA-seq) data yields higher quality data with less noise and larger dynamic range [32].

We therefore compiled PRECISE, a Precision RNA-seq Expression Compendium for Independent Signal Exploration. This high-fidelity expression profile compendium (median $R^2 = 0.98$ between biological replicates, see Figure A.1a,b) comprises 278 RNA-seq expression profiles across 154 unique experimental conditions for *Escherichia coli* K-12 MG1655 and BW25113. To assemble PRECISE, we collected and processed RNA-seq data from over 15 studies published by our research group (See Methods), comprising approximately 20% of all publicly available RNA-seq data in NCBI GEO [33] for *E. coli* K-12 MG1655 and BW25113 (Figure A.1c). The datasets

in PRECISE were generated in a single laboratory and obtained using a standardized protocol, with detailed reporting of experimental conditions and metadata to assist in usage as a comprehensive resource. This homogeneity mitigated batch effects (Figure A.1d,e), simplifying the data processing pipeline (See Methods).

We applied ICA to identify independent sources of variation in gene expression in PRECISE. The traditional use of ICA as a signal decomposition algorithm is illustrated in Figure 2.1a. When applied to transcriptomics data, ICA decomposes a collection of expression profiles ($\mathbf{X}$) into (1) a set of components, which represent underlying biological signals ($\mathbf{S}$), and (2) the components' condition-specific activities ($\mathbf{A}$) (Figure 2.1b,c). Each component, represented by a column of $\mathbf{S}$, contains a coefficient for each gene that represents the effect of a particular underlying signal on the gene's expression level. Components do not contain information on the condition-specific transcriptomic state. Conversely, ICA computes activity levels for each component across every condition in the compendium, represented by a row of $\mathbf{A}$, to account for condition-dependent expression changes. Each expression profile is represented by the summation over all components, each scaled by its condition-specific activity (Figure 1.3).

ICA of PRECISE produced 92 robust components that explained 68% of the expression variation (Figure A.1f). To make sense of each component, we applied the workflow described in Figure 2.2a. Most gene coefficients in a component were near zero, indicating that each underlying signal affects a small number of significant genes (Figure 2.2b). We removed genes with coefficients below a threshold (See Methods), resulting in a set of significant genes for each component. We defined these condition-invariant sets of genes as **i-modulons**, since these genes were **independently modulated** at consistent ratios across every condition in the compendium.

Henceforth, we will focus our discussion on the i-modulons extracted from the independent

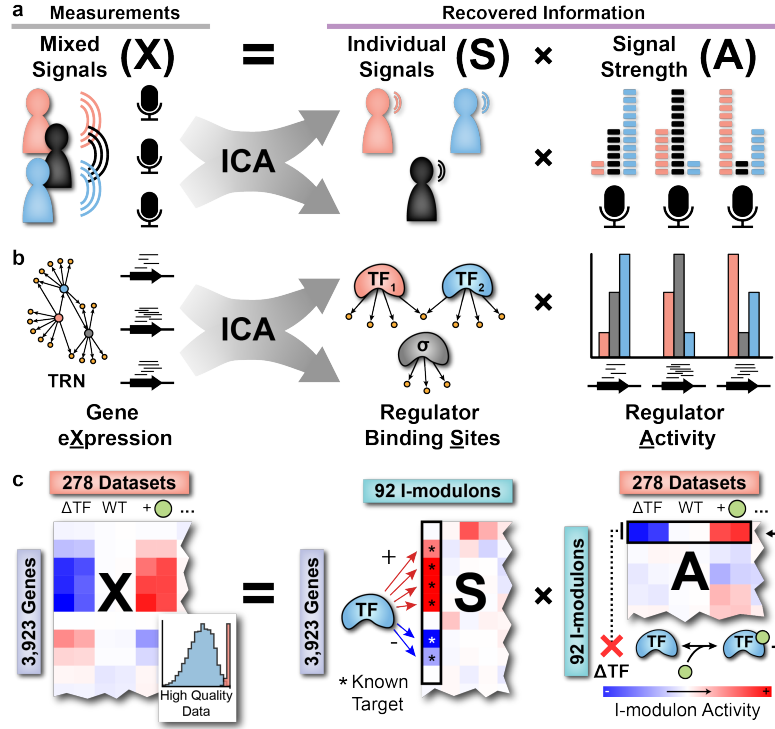


Figure 2.1: Independent component analysis and blind source separation. (a) Given three microphones recording three people speaking simultaneously, each microphone records each voice (i.e. signal) at different volumes (i.e. signal strengths) based on their relative distances. Using only these measured mixed signals, ICA recovers the original signals and their relative signal strengths by maximizing the statistical independence of the recovered signals [13, 34]. The mixed signals ($\mathbf{X}$) are a linear combination of the matrix of recovered source signals ($\mathbf{S}$) and the mixing matrix ($\mathbf{A}$) that represents the relative strength of each source signal in the mixed output signals. This relationship is mathematically described as $\mathbf{X} = \mathbf{S}\mathbf{A}$. (b) An expression profile under a specific condition can be likened to a microphone in a cell, measuring the combined effects of all transcriptional regulators. (c) Schematic illustration of ICA applied to a gene expression compendium. See A.1a,b for additional details on data quality. The example TF is a dual regulator that primarily up-regulates genes, and is activated by the green circular metabolite. Example experimental conditions shown are a TF knock-out, wild-type, and wild-type grown on medium supplemented with the activating metabolite. Each column of $\mathbf{X}$ contains an individual expression profile across 3,923 genes in *E. coli*.

components, and their respective activity levels as computed from ICA. We note that a gene may appear in multiple i-modulons if its expression is dependent on multiple underlying biological signals.

The 92 resulting i-modulons are listed in Table2.1. We hypothesized that each i-modulon

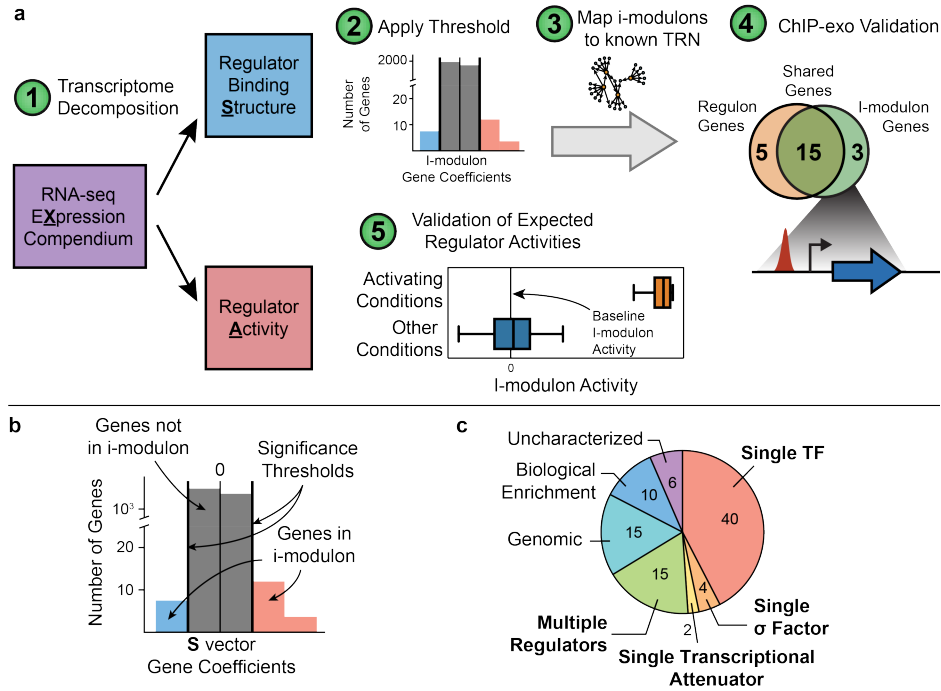


Figure 2.2: Characterization of independent components. (a) Schematic illustration of the workflow used to characterize independent components. (b) Independent components are processed into i-modulons by removing all genes with coefficients within a significance threshold (indicated in gray). Significant genes may have either positive (red) or negative (blue) coefficients. (c) Distribution of i-modulon categories. Categories of regulatory i-modulons are labelled in bold font. Genomic i-modulons account for single gene knock-outs, large deletions or duplications of genomic regions. Biological i-modulons contain genes enriched for a specific function, but are not linked to a specific transcriptional regulator.

was controlled by a particular transcriptional regulator, and that the i-modulon activity represented the condition-dependent activation state of the corresponding transcriptional regulator. To test this hypothesis, we examined the consistency between i-modulons and reported regulons, defined as the set of genes targeted by a common regulator, using a database of over 7,000 experimentally-derived regulatory interactions for *E. coli* [4].

Table 2.1: Overview of i-modulons extracted from PRECISE

I-modulon Name	# Genes	Regulator(s)	Precision	Recall	Biological Function
Amino Acid and Nucleotide Biosynthesis (9)					
ArgR	13	ArgR	0.92	0.1	Arginine biosynthesis
CysB	21	CysB	0.76	0.52	Sulfate assimilation
His-tRNA	9	His-tRNA	0.89	1	Histidine biosynthesis
Leu/Ile	14	IlvY or leu-tRNA or ile-tRNA	0.93	0.76	Branched-chain amino acid biosynthesis
Lrp	37	Lrp	0.86	0.16	Amino acid transport
MetJ	17	MetJ	0.65	0.73	Methionine biosynthesis
PurR-1	16	PurR	0.81	0.36	Purine biosynthesis
PurR-2	10	PurR	0.7	0.19	Pyrimidine biosynthesis
Tryptophan	11	TrpR or trp-tRNA or Tryptophan	0.82	0.5	Tryptophan biosynthesis
Carbon Source Utilization (17)					
AllR/AraC/FucR	18	AllR or AraC or FucR	1	0.6	Allantoin, glyoxylate, L-arabinose, and L-fucose catabolism
CdaR	10	CdaR	0.9	1	Glucarate catabolism
Cra	15	Cra	0.8	0.09	Glycolysis
Crp-2	40	Crp	0.7	0.05	Carbon transport
DhaR/Mlc	11	DhaR or Mlc	0.73	0.57	DHA kinase and PTS systems
ExuR/FucR	15	ExuR or FucR	0.67	0.63	D-galacturonate and L-fucose catabolism
GlcC	6	GlcC	1	0.86	Glycolate catabolism
GlpR	9	GlpR	1	1	Glycerol catabolism
GntR/TyrR	15	GntR or TyrR	0.87	0.54	Gluconate catabolism and tyrosine biosynthesis
MalT	9	MalT	0.89	0.8	Maltose catabolism
NagC/TyrR	17	NagC or TyrR	0.94	0.33	GlcNAC catabolism and tyrosine biosynthesis
PrpR	6	PrpR	0.67	0.8	Propionate catabolism

Table 2.1: Overview of i-modulons extracted from PRECISE, continued

I-modulon Name	# Genes	Regulator(s)	Precision	Recall	Biological Function
Pyruvate	8	BtsR or YpdB or PdhR	0.5	0.09	Pyruvate transport and metabolism
RbsR	6	RbsR	0.83	0.56	D-ribose catabolism
SrlR+GutM	5	SrlR and GutM	1	0.71	Sorbitol catabolism
XylR	13	XylR	0.46	0.86	Xylose catabolism
YiaJ	10	YiaJ	1	0.91	Ascorbate utilization
Energy Metabolism (4)					
ArcA-1	50	ArcA	0.66	0.07	Electron transport chain
ArcA-2	25	ArcA	0.84	0.05	Anaerobic response
Fnr	40	Fnr	0.85	0.08	Anaerobic response
NarL	29	NarL	0.93	0.23	Nitrate respiration
Metal Homeostasis (6)					
Copper	8	CusR or HprR or CueR	1	0.42	Copper homeostasis
FecI	10	FecI	0.5	1	Ferric citrate transport
Fur-1	48	Fur	0.9	0.25	Iron homeostasis
Fur-2	27	Fur	0.81	0.13	Iron homeostasis
NikR	6	NikR	0.83	0.83	Nickel homeostasis
Zinc	12	ZntR or Zur	0.58	1	Zinc homeostasis
Miscellaneous Metabolism (7)					
AtoC	6	AtoC	0.67	1	Acetoacetate degradation
Cbl+CysB	10	Cbl and CysB	0.8	0.89	Sulfur starvation
Crp-1	60	Crp	0.48	0.05	Miscellaneous Functions
CsqR	10	CsqR	0.5	1	Sulfoquinovose catabolism
FadR	13	FadR	0.69	0.41	Fatty acid degradation
GcvA	4	GcvA	0.75	0.75	Glycine cleavage system
PuuR	7	PuuR	1	1	Putrescine catabolism
Thiamine	11	Thiamine	1	1	Thiamine biosynthesis

Table 2.1: Overview of i-modulons extracted from PRECISE, continued

I-modulon Name	# Genes	Regulator(s)	Precision	Recall	Biological Function
Stress Response (11)					
CecR	5	CecR	1	1	Uncharacterized regulon
CpxR	17	CpxR	0.47	0.13	Membrane stress
EvgA	20	EvgA	0.5	0.63	Acid/osmotic stress
GadEWX	17	GadE and GadW and GadX	0.59	0.91	Acid stress response
GadWX	8	GadW and GadX	1	0.53	Acid stress response
Nac	37	Nac	0.86	0.06	Nitrogen limitation
NtrC+RpoN	56	NtrC and RpoN	0.57	0.67	Nitrogen limitation
OxyR	8	OxyR	1	0.17	Peroxide reductases
RpoH	13	RpoH	1	0.1	Heat shock response
RpoS	107	RpoS	0.37	0.13	General stress response
SoxS	41	SoxS	0.56	0.43	Oxidative stress response
Structural Components (3)					
FlhDC	41	FlhDC	0.93	0.49	Flagella assembly
FliA	30	FliA	0.97	0.45	Chemotaxis
RcsAB	28	RcsAB	0.36	0.63	Colanic acid capsule
Regulator Discovery (3)					
YgbI	9	YgbI			Unknown
YieP	11	YieP			Unknown
YneJ	5	YneJ			Unknown
Genomic Alterations (15)					
BW25113	17				See Section 2.2.6
crp-KO	14				Δcrp strains
deletion-1	41				39-gene deletion
deletion-2	110				171-gene deletion
duplication-1	84				129-gene deletion
e14-deletion	11				Removal of e14 prophage
efeU-repair	8				Accounts for <i>efeU</i> repair

Table 2.1: Overview of i-modulons extracted from PRECISE, continued

I-modulon Name	# Genes	Regulator(s)	Precision	Recall	Biological Function
entC-menF-KO	3				$\Delta entC$ and $\Delta menF$ strains
fur-KO	13				Δfur strains
gadWX-KO	2				$\Delta gadW$ and $\Delta gadX$ strains
insertion	9				IS2 overexpression
purR-KO	3				$\Delta purR$ strains
thrA-KO	3				$\Delta thrA$ strains
ydcI-KO	3				$\Delta ydcI$ strains
yheO-KO	16				$\Delta yheO$ strains
Biological Enrichment (10)					
curli	4				Curli assembly
fimbriae	8				Fimbriae assembly
flu-yeeRS	8				Genes in CP4-44 prophage
iron-related	20				Related to iron metabolism
lipopolysaccharide	36				LPS biosynthesis
membrane	15				Membrane proteins
nitrate-related	14				Nitric oxide response
proVWX	4				Glycine betaine transport
sgrT	4				Single gene <i>sgrT</i>
translation	24				Ribosomal subunits
Uncharacterized (7)					

We identified significant overlaps between regulons and 61 of the 92 i-modulons. We defined these 61 i-modulons as regulatory i-modulons. Four regulatory i-modulons were linked to a single sigma factor (RpoS, RpoH, FliA, and FecI), two i-modulons were linked to a single transcriptional attenuator (including the thiamine riboswitch) [35–37], and 40 i-modulons were linked to a single TF. Fifteen regulatory i-modulons were associated with multiple transcriptional regulators. Of the 31 non-regulatory i-modulons, 15 i-modulons were associated with distinct genetic changes, such as gene knock-outs and strain-specific differences, and 10 i-modulons were

enriched in a specific biological function or process (Figure 2.2c).

I-modulons were manually curated, and named after their associated regulator (e.g., the MetJ i-modulon) or biological function (e.g., the Tryptophan i-modulon). The majority of regulatory i-modulons (43 of 61) mapped to metabolic pathways (Figure 2.3), as previously noted [18]. The remaining regulatory i-modulons represented diverse cellular responses (Figure A.2b,c). Detailed information for all 92 i-modulons, including gene composition, regulon enrichments, activity levels, and upstream regulator binding motifs, is available in Appendix B. The PRECISE dataset and code for i-modulon computation are available on Github at <https://github.com/SBRG/precise-db>.

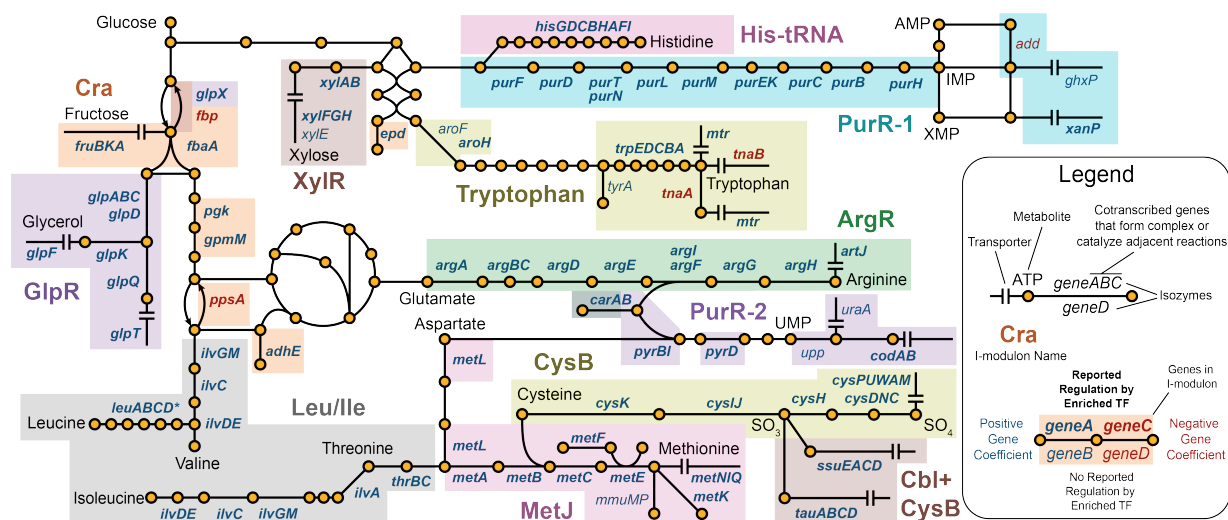


Figure 2.3: I-modulons are enriched with metabolic pathways. Twelve regulatory i-modulons are shown overlaying a metabolic network map. Gene names within a colored box are members of the corresponding i-modulon. Co-transcribed genes whose products form a complex or catalyze adjacent reactions share gene names (e.g., *trpEDCBA*), and isozymes (i.e., enzymes that catalyze identical reactions) share reaction lines (e.g., *glpABC/glpD*). Genes with positive i-modulon weights are colored blue, and genes with negative weights are red. Bold font indicates that the genes belong to the associated regulon.

2.2.2 Validation of i-modulon-regulator relationships

On average, 78% of genes in a regulatory i-modulon were reported targets of the linked transcriptional regulator(s) (Figure 2.4a,b). In order to benchmark the precision of the i-modulons generated from PRECISE, we applied the ICA workflow to ten randomly generated subsets of a single-platform microarray compendium [10], maintaining a similar number of experimental conditions as PRECISE (See Methods). In addition, we applied the ICA workflow to a microarray dataset generated by a single research group [26] that contained a similar number of samples as PRECISE. Both the proportion and precision of regulatory i-modulons were significantly lower (Mann-Whitney U test, p -value < 0.05) in the microarray datasets (Figure 2.4c,d). We also applied sparse-PCA to PRECISE and the dataset from a single research group, and found significantly less overlap with the known TRN (Mann-Whitney U test, p -value < 0.05), illustrating that independence, rather than sparsity, was key to obtaining regulatory i-modulons (Figure A.3a,b).

Many regulatory i-modulons contained genes that were not in the associated regulon. We hypothesized that these remaining genes were actually regulated by the associated regulator, but the binding sites were not experimentally determined. We tested this claim by performing ChIP-exo [38] to locate binding sites for the TFs MetJ and CysB, which regulate methionine biosynthesis and sulfate assimilation, respectively. We identified MetJ binding sites upstream of all 17 genes in the MetJ i-modulon (Figure 2.4e). The CysB regulon was split into the CysB i-modulon and the jointly regulated Cbl+CysB i-modulon (Figure A.3c). We identified CysB binding sites upstream of all genes except one (*iraD*) in both i-modulons (Figure 2.4f). TF binding was not detected near 9 of the 11 genes that were in the reported MetJ and CysB regulons but not in their respective i-modulons, potentially indicating inconsistencies in previous

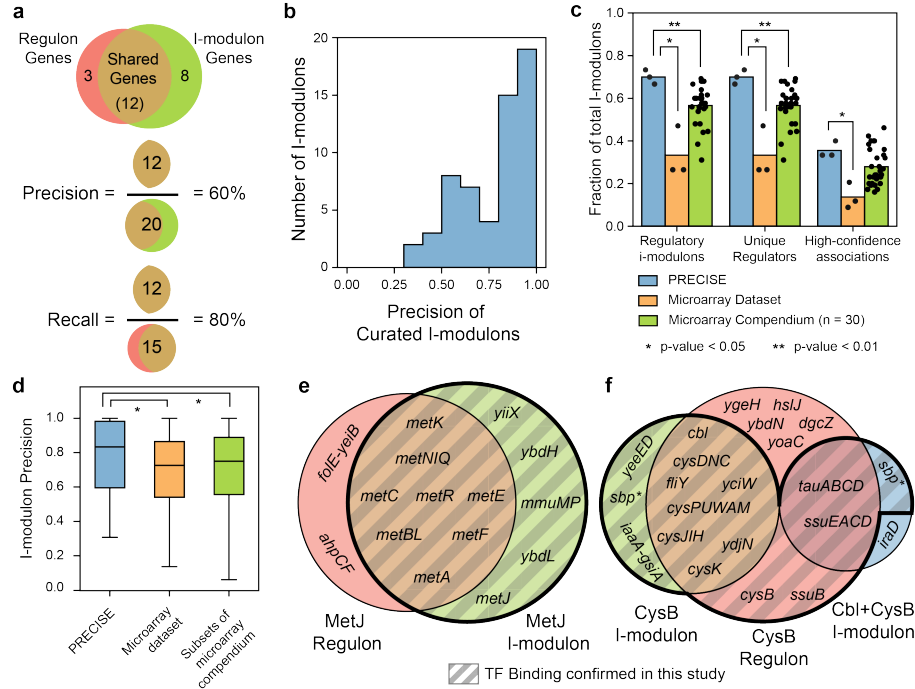


Figure 2.4: Validation of I-modulon-Regulator Relationships. (a) Precision is the fraction of genes in an i-modulon that are in the linked regulon, and recall is the fraction of genes in a regulon that are in the linked i-modulon. (b) Precision across all 61 regulatory i-modulons. (c) Fraction of total i-modulons significantly enriched with targets from a single transcriptional regulator. I-modulons generated from PRECISE (blue) were compared against i-modulons generated from a microarray dataset with 266 expression profiles [26] (orange), and i-modulons generated from 10 similar-sized subsets of a single-platform microarray compendium [10] (green). Each dataset was analyzed using 3-fold cross-validation (See Methods), resulting in 30 data points for the microarray compendium. Single star represents a Mann-Whitney U test p -value < .05, and double star represents a p -value < .01 (d) Boxplots comparing the precision of regulatory i-modulons across all three datasets. Single star represents a Mann-Whitney U test p -value < .05. Boxplot whiskers represent extrema of data, box bounds represent upper and lower quartiles, and center-line represents the median value. (e) Comparison of genes in the MetJ regulon (red) and i-modulon (green). Genes validated by ChIP-exo are shown in bold. Gene names for co-transcribed genes were combined (e.g. *metBL* represents the transcription unit containing *metB* and *metL*) (f) Comparison of genes in the CysB regulon (red) and the CysB and Cbl+CysB i-modulons (green and blue, respectively). Most genes in the Cbl+CysB i-modulon were regulated by both Cbl and CysB. The starred gene, *sbp*, was a member of both i-modulons but was not in the reported CysB regulon. Genes with TF binding as determined by ChIP-exo are shown in bold.

regulon definitions. The results from these two ChIP-exo experiments showed that previously unidentified regulator binding sites may explain the absence of genes in low-precision i-modulons.

Application of ICA also extracted the condition-specific activities of i-modulons, providing an additional source of validation. I-modulon activities were centered such that all i-modulon activities were zero for a reference condition (See Methods). Thus, i-modulon activities under a particular condition represented the relative up- or down-regulation of the i-modulon genes compared to the reference condition. It is important to note that positive i-modulon activities represent increased binding of transcriptional activators, or de-repression for transcriptional regulators.

In order to validate that media perturbations predictably altered specific i-modulon activities, we designed 10 expression profiling experiments to conditionally activate 20 regulators. We confirmed 75% (15/20) of predicted activations through 13 i-modulons (Figure 2.5). The sign of the i-modulon activity revealed whether the effector resulted in a net activation or repression of i-modulon genes, which was consistent with known mechanisms for the regulators.

Seven of the ten experiments included dual perturbations to simultaneously activate two regulators. In two cases, the two regulatory effects were recognized as a single signal, resulting in the combined i-modulons NagC/TyrR and GntR/TyrR (Figure A.3d,e). Combined i-modulons may also occur if a single molecule activates multiple regulators (Figure A.3f,g). Cytidine supplementation did not activate the cytidine-binding transcription factor CytR; however, the pyrimidine biosynthesis PurR-2 i-modulon was activated. Although four media additions (arginine, cytidine, galactose, and uracil) did not activate regulon-enriched i-modulons over the reference condition, the i-modulon structure of the TRN proved robust to additional data and displayed predictive capabilities.

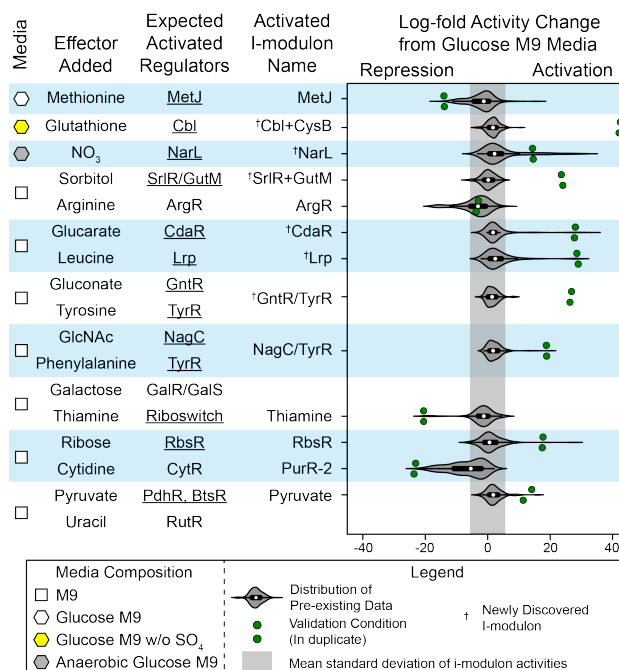


Figure 2.5: Validation of i-modulon activities. Compendium-wide activities of i-modulons predicted to be activated on ten media conditions. Correctly activated i-modulons are shown in bold. Distribution of i-modulon activities from pre-existing data includes all data from PRECISE excluding the ten validation conditions. The gray shaded region represents the average standard deviation across pre-existing i-modulon activities. All amino acid supplements were L-form, and all sugars were D-form. Abbreviations: GlcNAc, N-acetyl-glucosamine.

2.2.3 ICA reveals independent modules within the PurR regulon

In order to gain a detailed understanding of the biological roles of individual i-modulons, we programmatically generated a summary of characteristics for each i-modulon (See Appendix B). Figure 2.6 demonstrates the biological understanding to be gained from these characteristics for two exemplary i-modulons with significant overlap with genes in the PurR regulon (Fisher’s exact test p -value $< 10^{-10}$), named PurR-1 and PurR-2, respectively.

PurR is a repressor of nucleotide biosynthetic genes and is activated by intracellular purine levels [39]. The PurR-1 i-modulon contained 16 significant genes with both positive and negative coefficients, of which 14 were related to purine metabolism (See Figure 2.3). The PurR-

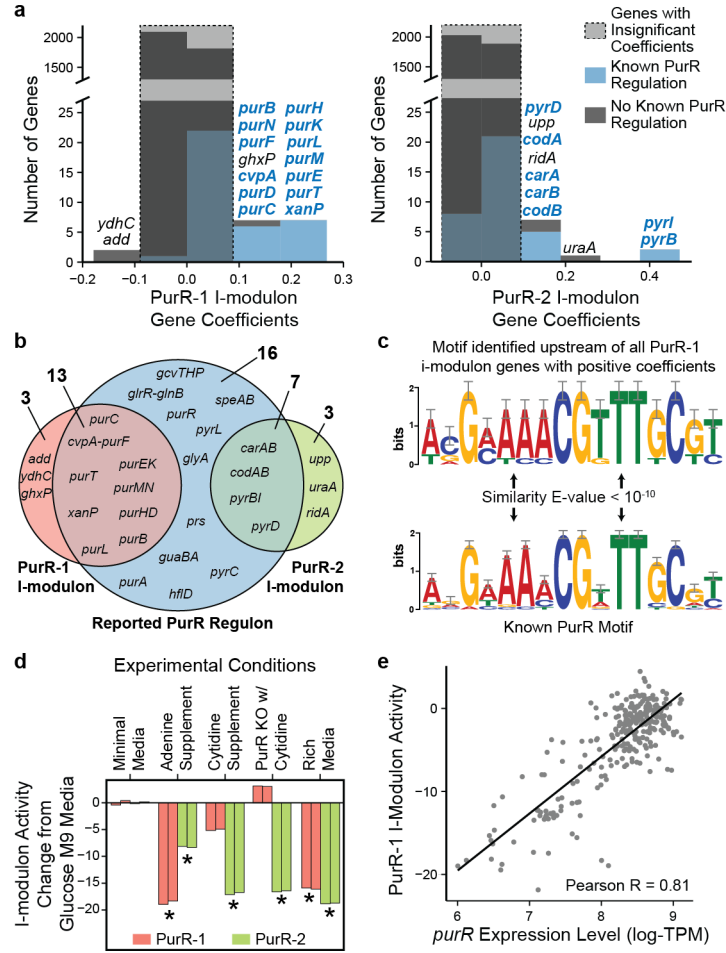


Figure 2.6: ICA reveals independent modules within the PurR regulon. (a) Histograms of gene coefficients in the PurR-1 and PurR-2 i-modulons. (b) Comparison of genes in the reported PurR regulon (blue), PurR-1 i-modulon (red) and PurR-2 i-modulon (green). Gene names for co-transcribed genes were combined (e.g. *codAB* represents *codA* and *codB*). (c) Motif identified upstream of genes in PurR-1 i-modulon compared to the reported PurR motif from RegulonDB [4]. This motif was identified upstream of the guanine/hypoxanthine transporter encoding gene *ghxP*, although regulator binding was not previously reported. (d) The two PurR associated i-modulons exhibited distinct responses to environmental perturbations. Asterisks denote significant i-modulon activities as compared to the reference condition (See Methods). Each bar represents a single biological replicate. (e) The PurR-1 i-modulon activity level is highly correlated with *purR* expression level across all conditions (excluding the PurR knock-out), whereas the PurR-2 i-modulon activity exhibits poor correlation (Figure A.4d). Similar information on all 92 i-modulons is available in Appendix B. Abbreviations: log-TPM, log-transformed Transcripts per Million.

2 i-modulon contained 10 genes, of which nine were in the pyrimidine biosynthetic pathway (Figure 2.6a). Together, the two PurR-related i-modulons accounted for 20 of the 36 genes in

the reported PurR regulon (Figure 2.6b). Segmentation of regulons into multiple constituent i-modulons was observed for other global regulators such as Fur and Crp (Figure A.4a,b).

The 14 genes with positive coefficients in the PurR-1 i-modulon were associated with purine biosynthesis, with 13 genes confirmed to be regulated by PurR. We detected the PurR binding motif upstream of the missing gene (*ghxP*), suggesting that it is regulated by PurR (Figure 2.6c). Similar analysis identified 150 previously unidentified regulator binding sites across 20 regulatory i-modulons. The two genes with negative coefficients (*add* and *ydhC*) were expressed inversely to the activation of the purine biosynthetic pathway; *add* encodes the first enzyme in the purine degradation pathway, and *ydhC* is a putative transporter. Since the negative gene coefficient indicated that *ydhC* expression responded inversely to purine biosynthetic gene expression, we hypothesized that *ydhC* was a purine-related efflux pump. In a similar fashion, we used the i-modulon structure to generate additional information for 224 genes with poor annotations, including 11 transporters. One such prediction, *yjiY*, was recently independently verified as a pyruvate transporter and renamed to *btsT* [40].

The condition-specific activities of the PurR-related i-modulons revealed differences in purine and pyrimidine biosynthetic gene expression (Figure 2.6d). Adenine supplementation activated the repressor PurR to decrease the activity of the PurR-1 and PurR-2 i-modulons, whereas cytidine supplementation resulted in a decrease in PurR-2 activity. Knock-out of PurR did not affect the PurR-2 activity, while de-repressing the purine biosynthetic pathway. LB rich medium decreased both i-modulon activities.

The relationship between the quantitative activities of the PurR i-modulons and the expression level of PurR indicated the drivers of regulator activity. The activity of the PurR-1 i-modulon was highly correlated (Pearson $R = 0.81$, $p\text{-value} < 10^{-10}$) with the expression level

of *purR* (Figure 2.6e). A similar relationship was observed in 26 of 61 regulatory i-modulons, 20 of which required a minimum expression level to activate the i-modulon (See Figure A.4c). In contrast, the activity of the PurR-2 i-modulon was poorly correlated with *purR* expression (Pearson $R = 0.3$, $p\text{-value} = 6 \times 10^{-7}$), and is likely controlled by UTP-dependent reiterative transcription [41] (Figure A.4d). The low correlation between many i-modulon activities and their corresponding regulator expression levels specifically identifies where static TRNs are inconsistent with expression data [42].

The results presented in this section demonstrate that the i-modulon structure of the TRN revealed by ICA provides a deep understanding of its biological functions and offers a guide to discovery.

2.2.4 ICA provides answers to unasked questions

Most expression profiles in PRECISE were designed to test specific biological hypotheses leading to independent, self-contained studies. However, by integrating these datasets together and applying ICA, we gained the ability to answer questions not addressed in the individual studies.

To illustrate this concept, we examined the i-modulons activated by the addition of 80 new expression profiles from four recently-published RNA-seq datasets [43–46] (Figure 2.7a). The expression profiles were generated and processed using the same protocol as the rest of the compendium. ICA of the original 198 expression profiles resulted in 73 i-modulons, 65 of which were maintained upon incorporation of the 80 additional expression profiles. Analysis of the i-modulons resulting from three nested subsets of PRECISE confirmed that i-modulons were rarely lost upon addition of new data (See Methods, Figure 2.7b). The 80 additional datasets

activated 26 new i-modulons, including 14 regulatory i-modulons and 9 genomic i-modulons. The activities of each new i-modulon revealed the causative expression profiles (Figure 2.7c).

For example, one dataset included knock-outs of 10 uncharacterized TFs [43], which generated five new i-modulons. Two i-modulons were dominated by a single knocked-out gene (Figure A.4e), whereas two remaining i-modulons (YiaJ and YieP) primarily contained genes with experimentally determined TF binding sites (Figure 2.7d,e), revealing two new high-confidence regulons. Knock-out of *cecR* produced an i-modulon that was recently independently validated with high accuracy [47]. Since the genes in the three new regulatory i-modulons had positive coefficients and the i-modulon activity was positive upon gene knock-out (Figure 2.7c), we concluded that the TF knock-out resulted in overexpression of these genes. Therefore, we characterized these TFs as transcriptional repressors.

The other three datasets contained endpoint strains from adaptive laboratory evolutions (ALEs) [44–46]. ALE endpoint strains often contain many mutations whose downstream effects are difficult to resolve [48]. However, most strains contained at least one mutation in or near a transcriptional regulator that corresponded with explainable changes in i-modulon activities (Figure 2.7c). Most of these regulators were well-characterized repressors and resulted in derepression of the i-modulon genes.

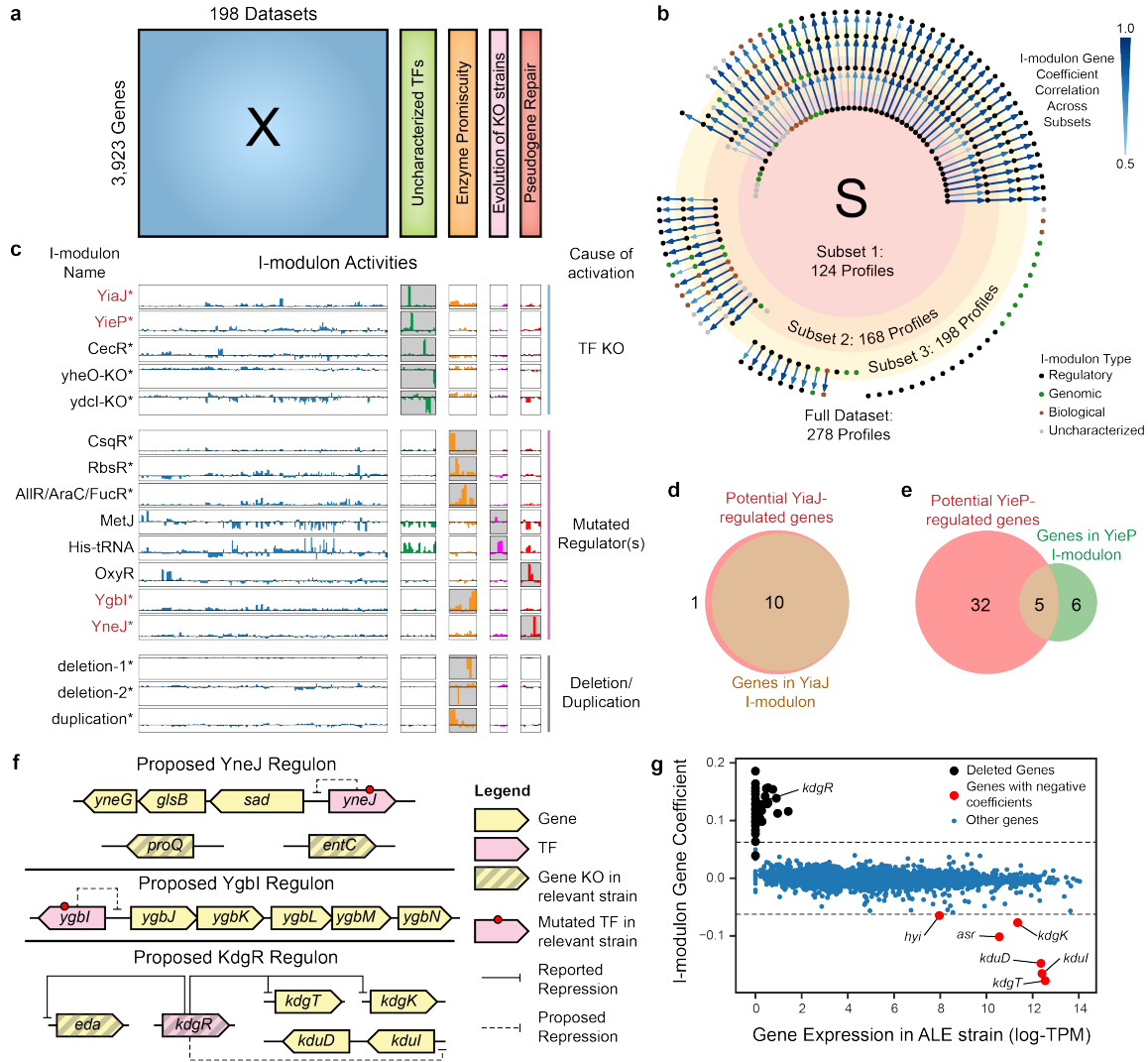


Figure 2.7: ICA provides answers to unasked questions. (a) Schematic illustration of appending four new datasets [43–46] to PRECISE. (b) Comparison of ICA results on three nested subsets of the PRECISE compendium. Each node represents an i-modulon and is colored by type (e.g. regulatory, genomic). Components are linked by an arrow if their gene coefficients are correlated (Pearson $R > 0.5$). Arrow widths and color represent correlation strength. (c) Compendium-wide activities for selected i-modulons. Each bar represents the activity of the denoted i-modulon in a single expression profile. I-modulons in bold were identified after adding the four new datasets. I-modulon in red propose regulons for previously uncharacterized TFs. I-modulons are grouped based on the genetic perturbation (e.g. TF KO, mutation) that activated the specific i-modulon. The dataset responsible for the i-modulon activation is highlighted in gray for each i-modulon. (d) Venn diagram comparing the YiaJ i-modulon to ChIP-exo-determined binding sites for YiaJ. (e) Venn diagram comparing the YieP i-modulon to ChIP-exo-determined binding sites for YieP. (f) Predicted regulatory roles for YneJ, YgbI and KdgR. (g) Scatter plot of gene expression in strain with 39-gene deletion against Deletion-1 i-modulon gene coefficients.

However, two i-modulons were activated in strains where mutations occurred within TFs with no published information (YgbI and YneJ). The YneJ i-modulon contained six genes, of which *proQ* and *entC* are knocked-out in the strain with highest i-modulon activities (Figure 2.7f). Three remaining genes were divergently transcribed from *yneJ*, hinting at a potential regulator binding site.

The YgbI i-modulon contained 9 genes, five of which are divergently transcribed from *ygbI*, and have putative functions for four-carbon sugar catabolism [49]. The YgbI i-modulon was activated during adaptation of *E. coli* MG1655 to growth on tartrate, a four-carbon sugar. Although *E. coli* does not support native growth on tartrate, the evolved strains were able to utilize tartrate as the primary carbon source. Of the remaining four genes, two may be explained by other mutations present in this endpoint. Therefore, we hypothesize that YgbI is a transcriptional repressor that binds between *ygbI* and *ygbJ* (Figure 2.7g).

ALE strains often contain deletions or duplications of large regions in the genome, which are detected as genomic i-modulons. The Deletion-1 i-modulon resulted from a 39-gene deletion and had a negative activity for the strain with the deletion. This indicated that genes with positive i-modulon coefficients are not expressed in this strain, whereas genes with negative i-modulon coefficients are over-expressed in this strain. The deletion itself included KdgR, a transcription factor known to repress *kdgT* and *kdgK* [50]. These genes, along with *asr*, *hyi*, *kduD*, and *kduI*, had negative i-modulon coefficients, indicating that they were over-expressed when KdgR was knocked out. Since *kduD* and *kduI* encode genes with similar metabolic functions to *kdgK* (Figure A.4f), we predict that they are also repressed by KdgR.

When new experimental data was added to the existing compendium, answers were revealed to questions unasked by the original studies. Next, we show that correlations between

i-modulon activity levels can do the same by revealing relationships between the cellular functions represented by the i-modulons.

2.2.5 Two i-modulons characterize the ‘Fear-Greed’ trade-off

Can the i-modulon decomposition be utilized to understand a major genetic perturbation of the transcriptome? Adaptive laboratory evolution of *E. coli* [51] revealed two distinct point mutations in RpoB, the RNA polymerase subunit β , that shift cellular resources towards growth-related functions (i.e. greed) away from stress-hedging functions (i.e. fear) [52].

The quantitative i-modulon activities for two RpoB mutant strains reflected this trade-off (Figure 2.8a). Six i-modulons whose activities significantly deviated from the wild-type strain were initially uncharacterized. Further investigation into one of these i-modulons revealed genes encoding translation machinery, such as ribosomal proteins, to comprise one of the uncharacterized i-modulons (Figure 2.8b). The compendium-wide activity of this Translation i-modulon was correlated with growth rate (Pearson $R = 0.59$, $p\text{-value} < 10^{-10}$, Figure A.5a), consistent with previous observations that growth is propelled by increased ribosomal catalytic activity [53–56]. The Translation i-modulon therefore represented the greedy, growth-related functions of the transcriptome.

The i-modulon with the largest activity decrease in both variants was enriched in genes controlled by the stress response sigma factor (RpoS). The RpoS i-modulon activity was correlated (Pearson $R = 0.65$, $p\text{-value} < 10^{-10}$, Figure A.5b) with the expression level of *rpoS* and revealed a quantitative measure of cellular stress across diverse conditions (Figure 2.8c). Therefore, the RpoS i-modulon represented the fearful stress-hedging functions of the transcriptome.

PRECISE also contained expression data for over 45 strains from ALE experiments [48],

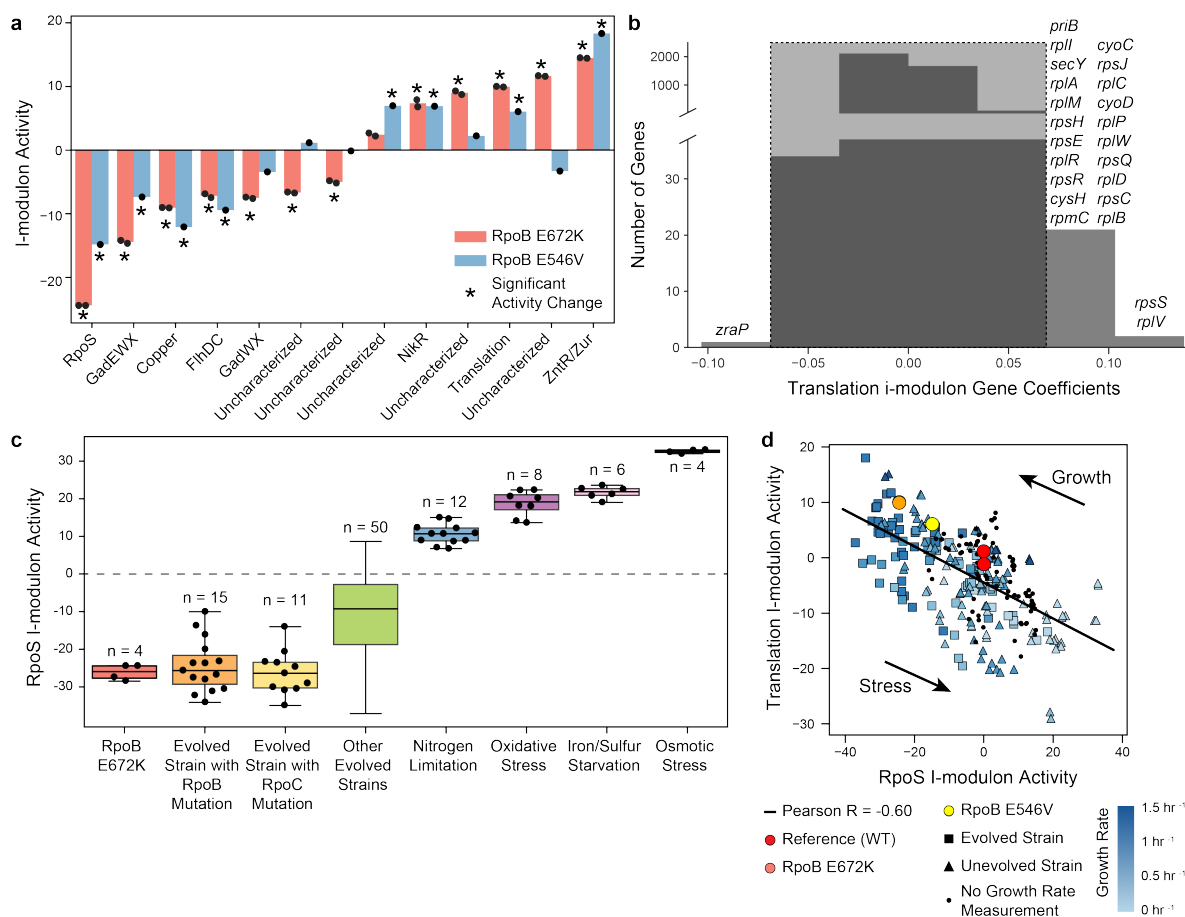


Figure 2.8: Two i-modulons characterize the ‘Fear vs. Greed’ trade-off. (a) Comparison of i-modulon activities in the RpoB E672K and RpoB E546V mutant strains grown on glucose minimal media against wild-type activities. Significant i-modulon activities are designated by asterisks (See Methods). For detailed information about these i-modulons, see Appendix B. (b) Histogram of translation i-modulon gene coefficients. Gene names are shown for genes above threshold. (c) The RpoS i-modulon activities revealed the stress level of the cell under various conditions. Box plot whiskers represent extrema, box bounds represent upper and lower quartiles, and center-line represents the median value. (d) The RpoS i-modulon activities were anti-correlated with the Translation i-modulon activities, highlighting the trade-off between stress-hedging and growth. Single nucleotide mutations in RpoB (in yellow and orange) shift cellular resources along this line from the wild-type strain (in red). Points were colored by growth rate measurements when available.

many of which contained mutations in genes encoding RNA polymerase subunits. All strains with mutated *rpoB* or *rpoC* genes exhibited low RpoS i-modulon activity, reflecting a reduction in stress-related expression. Further examination revealed that the RpoS i-modulon activity was

anti-correlated with the Translation i-modulon activity (Pearson $R = -0.60$, $p\text{-value} < 10^{-10}$), illuminating the compendium-wide transcriptomic trade-off between fear and greed (Figure 2.8d). The mutations in *rpoB* shift the strains along this line, increasing growth and reducing stress-related gene expression.

The results presented in this section show that the fear-greed trade-off, and other transcriptome restructuring events, can be studied in great detail by decomposing the transcriptome into a summation of independent regulatory events.

2.2.6 An i-modulon discriminates between two *E. coli* strains

The PRECISE compendium includes 46 RNA-seq datasets from *E. coli* BW25113, a closely-related strain to *E. coli* MG1655. The BW25113 strain was the background strain for the Keio collection of over 3,000 single-gene knock-outs [57]. The transcriptomic differences between these strains, resulting from 29 genetic variations [58], have not been characterized. We identified a single i-modulon whose activities separated the transcriptomes of the two strains (Figure 2.9a).

The i-modulon gene coefficients elucidated nine transcriptomic differences explained by genetic differences between the strains (Figure 2.9b). Deletions and truncation in the BW25113 strain account for all genes with negative coefficients. An insertion sequence (IS30) in the *mhpC* gene in the BW25113 strain corresponds to a large increase in expression of *mhpCDEF*, as IS30 contains a known promoter [59]. Point mutations at the predicted transcription start site of *tabA*, in the FabR regulator, and in the phenylalanine tRNA *pheV*, account for other genes with positive coefficients.

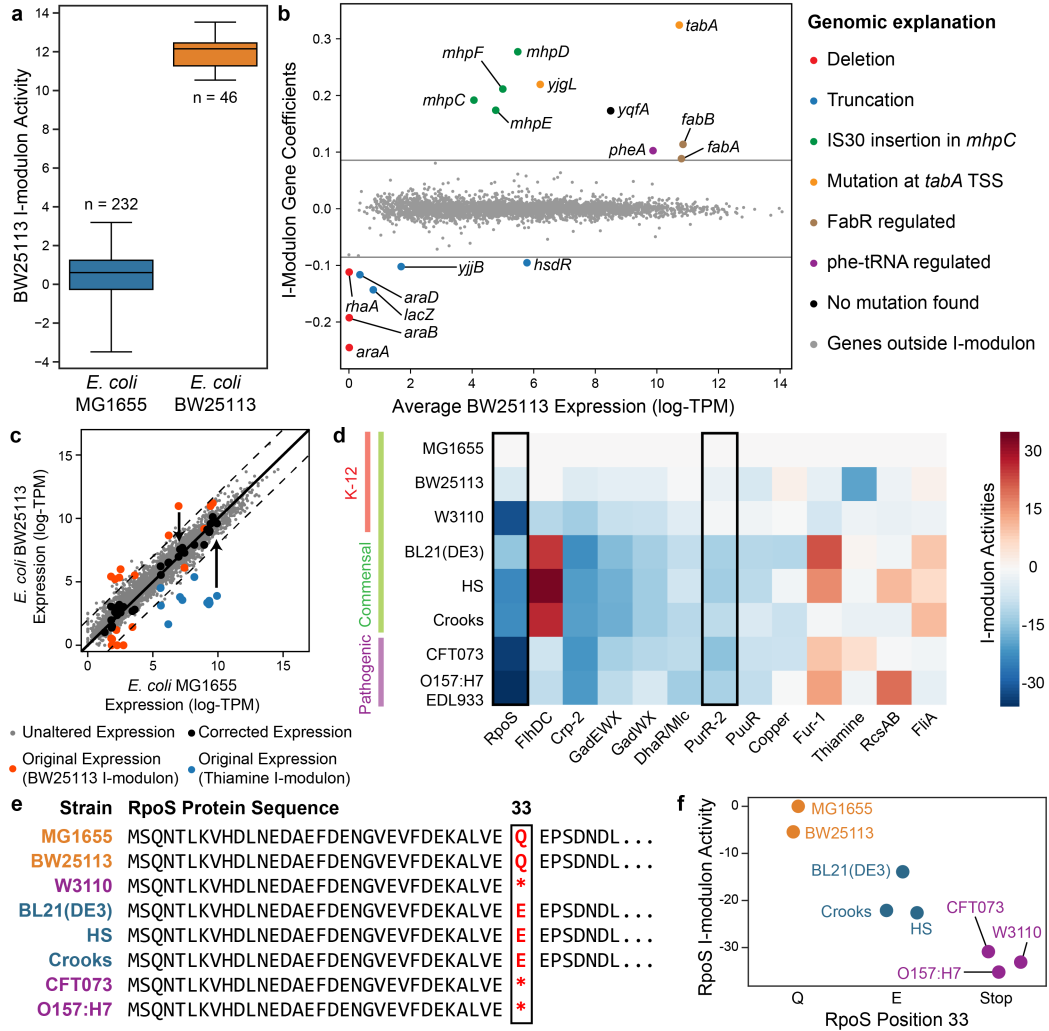


Figure 2.9: I-modulons explain expression changes across *E. coli* strains. (a) BW25113 i-modulon activities separated by strain. Number of expression profiles from each strain is shown. Boxplot whiskers represent extrema of data, box bounds represent upper and lower quartiles, and center-line represents the median value. (b) Scatterplot of average BW25113 expression against BW25113 i-modulon activity. (c) Subtraction of the BW25113 and Thiamine i-modulons from the *E. coli* BW25113 expression profile accounts for the major transcriptomic deviations from *E. coli* MG1655 grown without thiamine. Dashed lines indicate four-fold difference in TPM. (d) Heatmap of estimated i-modulon activities for 8 *E. coli* strains grown on glucose minimal media (with added thiamine and ferric chloride for BW25113). Only significantly altered regulatory i-modulon activities are shown. Boxed i-modulon activities are referred to in the main text. (e) Sequence alignment of the RpoS protein across the eight *E. coli* strains. (f) RpoS activities of the eight strains grouped by position 33 in the RpoS protein sequence, as detailed in panel (e). Abbreviations: TSS, transcription start site.

We then sought to use the summation of i-modulons to quantitatively account for these

strain-specific differences. To this end, we analyzed two expression profiles in the compendium: the reference condition (wild-type *E. coli* MG1655) and *E. coli* BW25113 with thiamine and ferric chloride supplementation. Only two i-modulons were differentially activated between the two conditions: the BW25113 i-modulon (described above) and the Thiamine i-modulon that controls thiamine biosynthesis through a riboswitch [35]. We accounted for the transcriptomic differences by subtracting the gene coefficients in the two i-modulons scaled by their respective i-modulon activities from the *E. coli* BW25113 expression profile. This subtraction increased the R^2 between the two strains' expression profiles from 0.29 to 0.96 for the BW25113 i-modulon genes, and from below 0 to 0.95 for the Thiamine i-modulon genes. (Figure 2.9c). A similar increase in the R^2 value for the BW25113 i-modulon genes was observed when the i-modulon was subtracted from any BW25113 expression profile. This illustrated that the summation of the independently modulated genes captured the major expression differences between the two strains.

In the previous section, we used the i-modulon structure to interpret the effects of a single mutation in *rpoB* on the transcriptome. Here, we illustrated the broader ability of the i-modulons to interpret and quantitatively account for the transcriptional differences between closely related strains. Next, we evaluate if i-modulons can provide understanding of larger scale genetic differences between strains.

2.2.7 I-modulons explain expression changes across *E. coli* strains

It has proven difficult to compare transcriptional regulation between different strains of the same species [60]. We examined the ability of i-modulons to provide a structured basis for such comparison. We grew and expression profiled a set of eight diverse *E. coli* strains in

identical media (with additional supplements for BW251123 as described above) and calculated their i-modulon activities using the 92 previously identified i-modulons as a basis (Figure 2.9d). Three strains (MG1655, BW25113 and W3110) diverged from the same ancestral K-12 strain with limited genetic differences; CFT0173 and O157:H7 EDL933 are pathogenic strains; and the remaining three strains (BL21(DE3), HS, and Crooks) are laboratory strains.

The expression profiles of the K-12 strains shared similar i-modulon activities, including higher activities in the pyrimidine-responsive PurR-2 i-modulon. This increased expression can be explained by a defect in the *rph-pyrE* operon, which leads to pyrimidine starvation in the K-12 strains [61]. The RpoS i-modulon activity was significantly suppressed in all strains except MG1655 and BW25113. Strains W3110, CFT073, and O157:H7 EDL933 have an amber mutation that results in a truncation of *rpoS* and is known to reduce its expression and activity [62, 63]. Three other strains (BL21(DE3), HS, and Crooks) contain a divergent mutation at the same position that likely explains their reduced RpoS i-modulon activity (Figure 2.9e,f).

These results demonstrate that the i-modulons derived from a single strain can provide a scaffold to analyze transcriptomes from other strains of the same species. Strain-specific differences in i-modulon activities can be traced to sequence variations in the associated regulators, thus providing a deep explanation for targeted strain differences.

2.3 Discussion

We have demonstrated that the combination of (1) independent component analysis of high-quality RNA-seq data and (2) high-resolution comprehensive regulator binding site information, identifies linear combinations of quantitative regulatory signals that reconstitute the *E. coli* transcriptome, leading to the first *E. coli* TRN inferred from an RNA-seq compendium. This

result suggests that a principle of i-modulon addition governs the composition of the *E. coli* transcriptome. Application of this principle provided a multidimensional understanding of the *E. coli* TRN, and uncovered detailed responses to environmental and genetic perturbations, optimality of adaptation to new conditions, and links between genotypes and phenotypes of *E. coli* strains. If this principle governs the composition of other prokaryotic transcriptomes, it provides a path to develop a detailed understanding of transcriptional regulation in less-understood organisms.

We have shown that for the model prokaryote *E. coli*, 61 of the 92 identified i-modulons represent the effects of characterized transcriptional regulators. This coverage is a consequence of the quality and diversity of the RNA-seq compendium used, the extensive information available on *E. coli* transcriptional regulators, and the relative simplicity of the *E. coli* TRN. In principle, if we obtained expression data for every condition sensed by a prokaryote, we could decompose its expression state to a non-reducible set of regulatory signals. These fundamental signals, combined with high-throughput binding data for all TFs in an organism [43], would lead to the establishment of a comprehensive quantitative transcriptional regulatory network.

2.4 Methods

2.4.1 RNA extraction and library preparation

Total RNA was sampled from duplicate cultures. All strains were grown in minimal salts (M9) medium at exponential phase, with complete growth conditions listed in the <https://github.com/SBRG/precise-db> repository. Growth curve analysis were performed using Bioscreen C Reader system with 200 uL culture volume per well. Two biological replicates were used in the assay. Media components were purchased from Sigma-Aldrich (St. Louis, MO). For

nitrate respiration cultures, a 35:50 ratio of carbon dioxide to nitrogen was bubbled through the media to deoxygenate. After inoculation and growth, 3 mL of cell broth ($OD_{600} \sim 0.5$) was immediately added to 2 volumes Qiagen RNA-protect Bacteria Reagent (6 mL), vortexed for 5 seconds, incubated at room temperature for 5 min, and immediately centrifuged for 10 min at 11,000 xg. The supernatant was decanted, and the cell pellet was stored in the -80°C . Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37°C . Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. The rRNA was removed using Illumina Ribo-Zero rRNA removal kit 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 300 bp. Libraries were ran on a HiSeq4000 (Illumina). All RNA-seq experiments were performed in biological duplicates from distinct samples.

2.4.2 ChIP-exo preparation

To activate each TF, cells were grown on relevant media: M9 minimal medium with 2 g/L glucose and 5 mM methionine for MetJ, M9 minimal medium with 2 g/L glucose and 0.25 mM taurine for CysB, and M9 minimal medium with 3.3 g/L pyruvate for YheO. To identify binding maps for each TF, DNA bound to each TF from formaldehyde cross-linked *E. coli* cells were isolated by chromatin immunoprecipitation (ChIP) with 15 ul specific antibodies that specifically recognize myc tag (9E10, Santa Cruz Biotechnology, Catalog # sc-40), and Dynabeads Pan Mouse IgG magnetic beads (Invitrogen) followed by stringent washings. ChIP materials

(chromatin-beads) were used to perform on-bead enzymatic reactions of the ChIP-exo method [38, 64, 65]. The sheared DNA of chromatin-beads was repaired by the NEBNext End Repair Module (New England Biolabs) followed by the addition of a single dA overhang and ligation of the first adaptor (5'-phosphorylated) using dA-Tailing Module (New England Biolabs) and NEBNext Quick Ligation Module (New England Biolabs), respectively. Nick repair was performed by using PreCR Repair Mix (New England Biolabs). Lambda exonuclease- and RecJf exonuclease-treated chromatin was eluted from the beads and overnight incubation at 65°C reversed the protein-DNA cross-link. RNAs- and Proteins-removed DNA samples were used to perform primer extension and second adaptor ligation with following modifications. The DNA samples incubated for primer extension as described previously were treated with dA-Tailing Module (New England Biolabs) and NEBNext Quick Ligation Module (New England Biolabs) for second adaptor ligation. The DNA sample purified by GeneRead Size Selection Kit (Qiagen) was enriched by polymerase chain reaction (PCR) using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The amplified DNA samples were purified again by GeneRead Size Selection Kit (Qiagen) and quantified using Qubit dsDNA HS Assay Kit (Life Technologies). Quality of the DNA sample was checked by running Agilent High Sensitivity DNA Kit using Agilent 2100 Bioanalyzer (Agilent) before sequenced using HiSeq 2500 (Illumina) following the manufacturer's instructions. ChIP-exo experiments were performed in biological duplicates from distinct samples.

2.4.3 ChIP-exo Processing

Sequence reads obtained from ChIP-exo experiments were mapped onto the *E. coli* reference genome (NC_000913.3) using bowtie (v1.1.2) [66] with default options in order to generate SAM output files. MACE program [67] was used to define peak candidates from biological dupli-

cates for each experimental condition with sequence depth normalization. Then, each peak was assigned to the nearest operon on either side, using operon definitions from RegulonDB. Only operons 500 base pairs downstream of peak were considered. Final operons on forward strand were required to be in front of the peak, and operons on reverse strand were required to be behind the peak. Genome-scale data were visualized using MetaScope to manually curate peaks (<https://sites.google.com/view/systemskimlab/software?authuser=0>).

2.4.4 Compilation of PRECISE

Raw sequencing reads were collected from GEO or produced using the above protocol, and mapped to the reference genome (NC_000913.3) using Bowtie (v1.1.2) [66] with the following options “-X 1000 -n 2 -3 3”. Transcript abundance was quantified using *summarizeOverlaps* from the R GenomicAlignments package (v1.18.0), with the following options “mode=”IntersectionStrict”, singleEnd=FALSE, ignore.strand=FALSE, preprocess.reads=invertStrand” [68]. To ensure the quality of the compendium, genes shorter than 100 nucleotides and genes with under 10 fragments per million mapped reads across all samples were removed before further analysis. Transcripts per Million (TPM) were calculated by DESeq2 (v1.22.1) [69]. The final expression compendium was log-transformed $\log_2(TPM + 1)$ before analysis, referred to as log-TPM. Biological replicates with $R^2 < 0.9$ between log-TPM were removed to reduce technical noise.

2.4.5 Compilation of the reported *E. coli* regulatory network

We compiled the global TRN using all interactions from RegulonDB 10.0 [3] for both transcription factor and sRNA binding sites. Binding sites were added from recent ChIP-exo studies [31], in addition to binding sites for Nac and NtrC [70] and potential binding sites for 10

uncharacterized transcription factors [43]. We also included sigma factor binding sites, riboswitch information, and transcriptional attenuation from Ecocyc [71]. When reported, mode of effect (i.e. activation or repression) was included. If the effect was unreported, or multiple effects were reported, effects were designated as unknown. All genes absent from PRECISE were removed from the final TRN.

2.4.6 Computing robust independent components

We first centered the compendium using wild-type *E. coli* MG1655 grown on glucose M9 minimal media as the reference condition (labelled *control_wt_glc_1* and *control_wt_glc_2*). We subtracted the mean expression of each gene in these two samples from the compendium to calculate log2-fold-change (LFC) deviations from the reference.

We used the Scikit-learn [72] (v0.19.2) implementation of the FastICA algorithm [34] to identify independent components. We executed FastICA 256 times with random seeds, a convergence tolerance of 10^{-8} , $\log(\cosh(x))$ as the contrast function, and the parallel search algorithm. We set the number of components in each iteration to the number of components that reconstruct 99% of the variance as calculated by principal component analysis (200 components).

The resulting source components (**S**) from each run were clustered using the Scikit-learn implementation of the DBSCAN algorithm [73], with epsilon of 0.1, and minimum cluster seed size of 128 samples (50% of the number of random restarts). DBSCAN does not require predetermination of the number of clusters, and does not require that all points belong to a cluster. The dimensionality of the dataset is therefore estimated by the number of clusters calculated by DBSCAN. The components computed by FastICA are standardized by default, with a mean of 0 and an L2-norm of 1. However, identical components from separate runs may

have opposite signs. Therefore we used the following distance metric:

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

where $\rho_{x,y}$ is the Pearson correlation between components x and y . Each component in a cluster was then inverted if necessary to ensure that the gene with the maximum absolute value in the component had a positive weight, creating sign-consistent clusters. The final independent components were defined as the centroid of each cluster in $\mathbf{S}$, and the weightings were defined as the centroid of their corresponding weighting vectors in $\mathbf{A}$.

In order to ensure that the final components were consistent across multiple runs, we computed the clustered components 100 times, and found that 92 components were identified in every run ($d_{x,y} < 0.1$ between components), which were the final robust components used in the analysis.

In order to confirm that the i-modulon structure was generally invariant to the composition of the expression compendium, we applied ICA to three subsets of PRECISE. The first subset consisted of data published before January 2018 from unevolved *E. coli* (124 profiles), the second subset consisted of all data published before January 2018 (168 profiles), and the third subset consisted of all data published before January 2018, plus data created to test the i-modulon structure (198 profiles). We compared the resulting components using the absolute value of the Pearson correlation coefficient. The resulting network was graphed using the Graphviz [74] python library (v0.9) (Figure 2.7b). Correlations below 0.5 were discarded as insignificant.

2.4.7 Determination of the gene coefficient threshold

Each component in $\mathbf{S}$ contains the contributions of each gene to the statistically independent source of variation. Most of these values are near zero for a given component. In order

to identify the most significant genes in each component, we modified the method proposed in Frigyesi et al. [75]. For each component, we iteratively removed genes with the largest absolute value and computed the D’Agostino K^2 test statistic for the resulting distribution. The D’Agostino K^2 statistic is a measure of the skew and kurtosis of a sample distribution [76]. Once the test statistic dropped below a cutoff, we designated the removed genes as significant.

To identify this cutoff, we performed a sensitivity analysis on the concordance between significant genes in each component and known regulons. First, we isolated the 20 genes from each component with the highest absolute gene coefficients. We then compared each gene set against all known regulons using the two-sided Fisher’s Exact Test ($\text{FDR} < 10^{-5}$). For each component with at least one significant enrichment, we selected the regulator with the lowest p-value.

Next, we varied the D’Agostino K^2 test statistic from 200 through 1000 in increments of 50, and computed the F1-score (harmonic average between precision and recall) between each component and its linked regulator. The maximum value of the average F1-score across the components with linked regulators occurred at a test statistic of cutoff of 550 (See Figure A.6a-c).

Since each set of significant genes represents a set of independently modulated genes, we henceforth refer to these gene sets as i-modulons. Since independent components have no canonical direction, we inverted i-modulons (and related activities) such that the number of positive genes in an i-modulon was always larger than the number of negative genes.

2.4.8 Associating regulators to i-modulons

We compared the set of significant genes in each i-modulon to each regulon (defined as the set of genes regulated by any given regulator) using the two-sided Fisher’s Exact Test ($\text{FDR} < 10^{-5}$). Additionally, combined regulon enrichments were calculated to identify joint regulation of i-modulons (such as NtrC+RpoN and NagC/TyrR), using both intersection (+) and union (/) of up to three regulons. Final i-modulon-regulator associations were determined through manual curation of enriched regulators.

2.4.9 Cumulative explained variance for ICA

Components were initially ordered by the L2-norm (sum of squares) of for each row in the $\mathbf{A}$ matrix for ICA. Cumulative explained variance was calculated for component K , as described in the EEGLAB suite [77]:

$$\text{CEV}(K) = 1 - \frac{\text{TSS}(\mathbf{X} - \sum_{k=0}^K \mathbf{s}_k \mathbf{a}_k)}{\text{TSS}(\mathbf{X})}$$

where

$$\text{TSS}(\mathbf{Y}) = \sum_{i,j} y_{i,j}^2$$

and $\text{TSS}(\mathbf{Y})$ is the total sum of squares, $\mathbf{X}$ is the original expression profile, $\mathbf{s}_k$ is column k in the $\mathbf{S}$ matrix, and $\mathbf{a}_k$ is row k in the $\mathbf{A}$ matrix.

2.4.10 Comparison of microarray data and PRECISE

We acquired the microarray compendium from the DREAM5 network inference challenge [10]. We removed expression profiles without biological replicates, and removed expression profiles with an R^2 score below 0.9 with its biological replicates. The final microarray compendium

contained 4289 genes and 461 expression profiles. We then randomly selected ten subsets of this compendium, each containing 154 unique conditions to mirror the composition of the PRECISE compendium. We included all experimental replicates of these conditions, resulting in ten datasets ranging between 255 to 289 total expression profiles. ICA was performed as described above for all ten datasets.

The microarray dataset from a single research group was acquired from NCBI GEO Series GSE6836 [26]. This dataset had similar size to PRECISE (266 experiments) to ensure comparability. Microarray data was processed using the RMA R package [78] (v1.50). ICA was performed as described above for both datasets. PCA determined that 148 components reconstructed 99% of the variance in the microarray dataset, and 104 robust independent components were identified.

Sparse-PCA was performed using the *elasticnet* R package [79] (v1.1.1), searching for the same number of components as with ICA (200 for PRECISE and 148 for microarray data), and a vector of ones as thresholding parameters.

For the comparison figures, each set of components was randomly split into three groups for 3-fold validation. For each training set, we selected the top 20 genes in each component and assigned the regulon with the lowest p-value from Fisher’s exact test ($\text{FDR} < 10^{-5}$). If no regulon contained significant p-values, the component was discarded. We then performed sensitivity analysis for the test statistic cutoff using these regulon assignments, searching for the test statistic value that maximized the F1-score across the training set.

The test statistic trained on the training set was then applied to the testing set to calculate the final validation i-modulons (i.e. significant genes). To assess the similarity between these i-modulons and known regulation, we identified the regulon with the lowest p-value ($\text{FDR} < 10^{-5}$) for each i-modulon in the test set. No manual curation was used to generate the comparison

figure.

2.4.11 Differential activity analysis

We first computed the distribution of differences in i-modulon activities between biological replicates, and then fit a log-normal distribution to each distribution. We confirmed that the difference in activities between biological replicates followed a log-normal distribution for all i-modulons using the Kolmogorov-Smirnov test and validating through quantile-quantile plots (Figure A.6d-f).

To test for differential activity of an i-modulon between two different conditions, we first computed the average activity of the i-modulon between biological replicates, if available. We then computed the absolute value of the difference in i-modulon activities between the two conditions. This difference was compared against the log-normal distribution for the i-modulon to calculate a p-value. I-modulons were designated as significant if the absolute value of their activities was greater than 5, and $FDR < 0.01$.

2.4.12 I-modulon summation

We selected samples *control__wt_glc__1* and *control__wt_glc__2* to represent the wild-type strain, and samples *omics__bw_glc__1* and *omics__bw_glc__2* to represent the mutated strains to be corrected. The average activities between replicates were used for the corrections. The corrections were applied to the BW25113 and Thiamine i-modulons.

The i-modulon decomposition is based on the equation

$$\mathbf{X} = \mathbf{S}\mathbf{A}$$

where

$$\mathbf{x}_j = \sum_i \mathbf{s}_i a_{i,j}$$

for a particular expression profile $\mathbf{x}_j$, where i represents an i-modulon. We aim to produce the correction ($\mathbf{x}'_2$) to the expression profile ($\mathbf{x}_2$) with respect to a reference expression profile ($\mathbf{x}_1$) for all differentially activated i-modulons $i \in I$:

$$\mathbf{x}'_2 = \mathbf{x}_2 - \sum \tilde{\mathbf{s}}_i (a_{i,2} - a_{i,1})$$

where $\tilde{\mathbf{s}}_i$ is a vector of zeros except for significant gene coefficients in i-modulon i , and $a_{i,j}$ is the activity of i-modulon i under condition j .

2.4.13 Multiple strain RNA-seq processing and projection

Raw sequencing reads and transcriptome abundance were identified similar to as described in the section above, using the following reference genomes: NC_000913.3 (MG1655 and BW25113), NC_007779.1 (W3110), NC_010468.1 (Crooks), NC_012971.2 (BL21(DE3)), NC_009800.1 (HS), NZ_CP008957.1 (O157:H7 EDL933), and NC_004431.1 (CFT073). Genes absent from a particular strain with respect to the reference strains (MG1655) were removed, leaving 915 core genes. We calculated the $\log_2(TPM + 1)$ values using the same centering to reference conditions (*control_wt_glc_1* and *control_wt_glc_2*) as described above.

Thereafter, we calculated the i-modulon activities for the eight new *E. coli* expression profiles using the previously identified 92 independent components (including all gene coefficients for the 915 conserved genes). We projected the eight new expression profiles ($\mathbf{X}'$) onto the previously computed basis ($\mathbf{S}$):

$$\mathbf{A}' = pinv(\mathbf{S})\mathbf{X}'$$

where $\mathbf{A}'$ represents the i-modulon activities for the eight strains, and $pinv$ is the pseudo-inverse function. This represents the least-squares approximation of $\mathbf{A}$.

Acknowledgements

We thank Dr. Joe Pogliano for biological insights, Dr. David Heckmann, Dr. James Yurkovich, Dr. Jared Broddrick, Erol Kavvas, Jean-Christophe Lachance, Saugat Poudel, Dr. Adam Feist and Dr. Ke Chen for helpful discussions and Marc Abrams for editorial comments. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was funded by the Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (grant number NNF10CC1016517).

Author Contributions: A.V.S. designed the study and performed the analysis, with assistance from L.Y. and B.O.P. Y.G., R.S., Y.H., and S.X. performed expression profiling experiments. Y.G. performed ChIP-exo experiments. D.K. performed expression profiling for multi-strain expression analysis, and K.S.C. analyzed the multi-strain data. L.Y. and Z.A.K. provided technical support and conceptual advice. A.V.S. and B.O.P. wrote the manuscript.

Chapter 2, in part, is a reprint of the material as appears in: **AV Sastry**, Y Gao, R Szubin, Y Hefner, S Xu, D Kim, KS Choudhary, L Yang, ZA King, and BO Palsson. 2019. “The *Escherichia coli* transcriptome mostly consists of independently regulated modules”. *Nature Communications*. 10 (5536). The dissertation author was the primary author.

2.5 References

1. Galagan, J. E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., Abeel, T., Mahwinney, C., Kennedy, A. D., Allard, R., Brabant, W., Krueger, A., Jaini, S., Honda, B., Yu, W.-H., Hickey, M. J., Zucker, J., Garay, C., Weiner, B., Sisk, P., Stolte, C., Winkler, J. K., Van de Peer, Y., Iazzetti, P., Camacho, D., Dreyfuss, J., Liu, Y., Dorhoi, A., Mollenkopf, H.-J., Drogaris, P., Lamontagne, J., Zhou, Y., Piquenot, J., Park, S. T., Raman, S., Kaufmann, S. H. E., Mohn, R. P., Chelsky, D., Moody, D. B., Sherman, D. R. & Schoolnik, G. K. The Mycobacterium tuberculosis regulatory network and hypoxia. en. *Nature* **499**, 178–183 (July 2013).
2. Buescher, J. M., Liebermeister, W., Jules, M., Uhr, M., Muntel, J., Botella, E., Hessling, B., Kleijn, R. J., Le Chat, L., Lecointe, F., Mäder, U., Nicolas, P., Piersma, S., Rügheimer, F., Becher, D., Bessieres, P., Bidnenko, E., Denham, E. L., Dervyn, E., Devine, K. M., Doherty, G., Drulhe, S., Felicori, L., Fogg, M. J., Goelzer, A., Hansen, A., Harwood, C. R., Hecker, M., Hubner, S., Hultschig, C., Jarmer, H., Klipp, E., Leduc, A., Lewis, P., Molina, F., Noirot, P., Peres, S., Pigeonneau, N., Pohl, S., Rasmussen, S., Rinn, B., Schaffer, M., Schnidder, J., Schwikowski, B., Van Dijl, J. M., Veiga, P., Walsh, S., Wilkinson, A. J., Stelling, J., Aymerich, S. & Sauer, U. Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. en. *Science* **335**, 1099–1103 (Mar. 2012).
3. Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeda, D., Muñoz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Del Moral-Chávez, V., Rinaldi, F. & Collado-Vides, J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. en. *Nucleic Acids Res.* **44**, D133–43 (Jan. 2016).
4. Santos-Zavaleta, A., Sánchez-Pérez, M., Salgado, H., Velázquez-Ramírez, D. A., Gama-Castro, S., Tierrafria, V. H., Busby, S. J. W., Aquino, P., Fang, X., Palsson, B. O., Galagan, J. E. & Collado-Vides, J. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. en. *BMC Biol.* **16**, 91 (Aug. 2018).
5. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. en. *Science* **306**, 636–640 (Oct. 2004).
6. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Fietze, S., Fu, Y., Gertz, J., Grubert, F., Harman, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J. J., Lian, J., Monahan, H., O’Geen, H., Ouyang, Z., Partridge, E. C., Patas, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapiro, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. & Snyder, M. Architecture of

- the human regulatory network derived from ENCODE data. en. *Nature* **489**, 91–100 (Sept. 2012).
7. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R. P. & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. en. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9186–9191 (May 2010).
 8. Lin, S. & Riggs, A. D. The general affinity of lac repressor for E. coli DNA: implications for gene regulation in procaryotes and eucaryotes. en. *Cell* **4**, 107–111 (Feb. 1975).
 9. Latchman, D. S. Transcription factors: an overview. en. *Int. J. Biochem. Cell Biol.* **29**, 1305–1312 (Dec. 1997).
 10. Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J. & Stolovitzky, G. Wisdom of crowds for robust gene network inference. en. *Nat. Methods* **9**, 796–804 (July 2012).
 11. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. en. *Nat. Rev. Microbiol.* **8**, 717–729 (Oct. 2010).
 12. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. en. *Nat. Commun.* **9**, 1090 (Mar. 2018).
 13. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36**, 287–314 (Apr. 1994).
 14. Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T. & Huang, X. A review of independent component analysis application to microarray gene expression data. en. *Biotechniques* **45**, 501–520 (Nov. 2008).
 15. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. en. *Bioinformatics* **18**, 51–60 (Jan. 2002).
 16. Chiappetta, P., Roubaud, M. C. & Torr sani, B. Blind source separation and the analysis of microarray data. en. *J. Comput. Biol.* **11**, 1090–1109 (2004).
 17. Martoglio, A.-M., Miskin, J. W., Smith, S. K. & MacKay, D. J. C. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. en. *Bioinformatics* **18**, 1617–1624 (Dec. 2002).
 18. Teschendorff, A. E., Journ e, M., Absil, P. A., Sepulchre, R. & Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. en. *PLoS Comput. Biol.* **3**, e161 (Aug. 2007).
 19. Engreitz, J. M., Daigle Jr, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. en. *J. Biomed. Inform.* **43**, 932–944 (Dec. 2010).

20. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. en. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10101–10106 (Aug. 2000).
21. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebreton, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A. & Radvanyi, F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. en. *Cell Rep.* **9**, 1235–1245 (Nov. 2014).
22. Huang, D.-S. & Zheng, C.-H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. en. *Bioinformatics* **22**, 1855–1862 (Aug. 2006).
23. Zhang, X. W., Yap, Y. L., Wei, D., Chen, F. & Danchin, A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. en. *Eur. J. Hum. Genet.* **13**, 1303–1311 (Dec. 2005).
24. Karczewski, K. J., Snyder, M., Altman, R. B. & Tatonetti, N. P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. en. *PLoS Genet.* **10**, e1004122 (Feb. 2014).
25. Moretto, M., Sonogo, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeda, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., Collado-Vides, J., Meysman, P. & Engelen, K. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. en. *Nucleic Acids Res.* **44**, D620–3 (Jan. 2016).
26. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. & Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. en. *PLoS Biol.* **5**, e8 (Jan. 2007).
27. Kim, M., Rai, N., Zorraquino, V. & Tagkopoulos, I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. en. *Nat. Commun.* **7**, 13090 (Oct. 2016).
28. Lewis, N. E., Cho, B.-K., Knight, E. M. & Palsson, B. O. Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. en. *J. Bacteriol.* **191**, 3437–3444 (June 2009).
29. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. en. *Nat. Biotechnol.* **32**, 896–902 (Sept. 2014).
30. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. & Irizarry, R. A. Tackling the widespread and critical impact of batch effects in high-throughput data. en. *Nat. Rev. Genet.* **11**, 733–739 (Oct. 2010).

31. Fang, X., Sastry, A., Mih, N., Kim, D., Tan, J., Yurkovich, J. T., Lloyd, C. J., Gao, Y., Yang, L. & Palsson, B. O. Global transcriptional regulatory network for *Escherichia coli* robustly connects gene expression to transcription factor activities. en. *Proc. Natl. Acad. Sci. U. S. A.* (Sept. 2017).
32. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. en. *Nat. Rev. Genet.* **10**, 57–63 (Jan. 2009).
33. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (Jan. 2013).
34. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. en. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999).
35. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. en. *Trends Biochem. Sci.* **29**, 11–17 (Jan. 2004).
36. Henkin, T. M. & Yanofsky, C. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. en. *Bioessays* **24**, 700–707 (Aug. 2002).
37. Kolter, R. & Yanofsky, C. Attenuation in amino acid biosynthetic operons. en. *Annu. Rev. Genet.* **16**, 113–134 (1982).
38. Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* **100**, 21–24 (2012).
39. Cho, B.-K., Federowicz, S. A., Embree, M., Park, Y.-S., Kim, D. & Palsson, B. Ø. The PurR regulon in *Escherichia coli* K-12 MG1655. en. *Nucleic Acids Res.* **39**, 6456–6464 (Aug. 2011).
40. Kristoficova, I., Vilhena, C., Behr, S. & Jung, K. BtsT - a novel and specific pyruvate/H⁺ symporter in *Escherichia coli*. en. *J. Bacteriol.* (Oct. 2017).
41. Turnbough Jr, C. L. & Switzer, R. L. Regulation of pyrimidine biosynthetic gene expression in bacteria: repression without repressors. en. *Microbiol. Mol. Biol. Rev.* **72**, 266–300, table of contents (June 2008).
42. Larsen, S. J., Röttger, R., Schmidt, H. H. H. W. & Baumbach, J. E. *coli* gene regulatory networks are inconsistent with gene expression data. en. *Nucleic Acids Res.* **47**, 85–92 (Jan. 2019).
43. Gao, Y., Yurkovich, J. T., Seo, S. W., Kabimoldayev, I., Dräger, A., Chen, K., Sastry, A. V., Fang, X., Mih, N., Yang, L., Eichner, J., Cho, B.-K., Kim, D. & Palsson, B. O. Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. en. *Nucleic Acids Res.* (Aug. 2018).

44. Guzmán, G. I., Olson, C. A., Hefner, Y., Phaneuf, P. V., Catoi, E., Crepaldi, L. B., Micas, L. G., Palsson, B. O. & Feist, A. M. Reframing gene essentiality in terms of adaptive flexibility. en. *BMC Syst. Biol.* **12**, 143 (Dec. 2018).
45. Anand, A., Olson, C. A., Yang, L., Sastry, A. V., Catoi, E., Choudhary, K. S., Phaneuf, P. V., Sandberg, T. E., Xu, S., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Pseudogene repair driven by selection pressure applied in experimental evolution. en. *Nat Microbiol* (Jan. 2019).
46. Guzmán, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, Á., Papp, H., de Raad, M., King, Z. A., Hefner, Y., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B. & Feist, A. M. Enzyme promiscuity shapes adaptation to novel growth substrates. en. *Mol. Syst. Biol.* **15**, e8462 (Apr. 2019).
47. Yamanaka, Y., Shimada, T., Yamamoto, K. & Ishihama, A. Transcription factor CecR (YbiH) regulates a set of genes affecting the sensitivity of Escherichia coli against cefoperazone and chloramphenicol. en. *Microbiology* **162**, 1253–1264 (July 2016).
48. Phaneuf, P. V., Gosting, D., Palsson, B. O. & Feist, A. M. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. en. *Nucleic Acids Res.* (Oct. 2018).
49. Zhang, X., Carter, M. S., Vetting, M. W., San Francisco, B., Zhao, S., Al-Obaidi, N. F., Solbiati, J. O., Thiaville, J. J., de Crécy-Lagard, V., Jacobson, M. P., Almo, S. C. & Gerlt, J. A. Assignment of function to a domain of unknown function: DUF1537 is a new kinase family in catabolic pathways for acid sugars. en. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4161–9 (July 2016).
50. Pouyssegur, J. & Stoeber, F. Genetic control of the 2-keto-3-deoxy-d-gluconate metabolism in Escherichia coli K-12: kdg regulon. en. *J. Bacteriol.* **117**, 641–651 (Feb. 1974).
51. LaCroix, R. A., Sandberg, T. E., O’Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Appl. Environ. Microbiol.* **81**, 17–30 (Jan. 2015).
52. Utrilla, J., O’Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. en. *Cell Syst* **2**, 260–271 (Apr. 2016).
53. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. en. *Science* **330**, 1099–1102 (Nov. 2010).
54. Scott, M., Klumpp, S., Mateescu, E. M. & Hwa, T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. en. *Mol. Syst. Biol.* **10**, 747 (Aug. 2014).

55. Valgepea, K., Adamberg, K., Seiman, A. & Vilu, R. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. en. *Mol. Biosyst.* **9**, 2344–2358 (Sept. 2013).
56. O’Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. en. *PLoS Comput. Biol.* **12**, e1004998 (June 2016).
57. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. en. *Mol. Syst. Biol.* **2**, 2006.0008 (Jan. 2006).
58. Grenier, F., Matteau, D., Baby, V. & Rodrigue, S. Complete Genome Sequence of Escherichia coli BW25113. en. *Genome Announc.* **2** (Oct. 2014).
59. Dalrymple, B. & Arber, W. Promotion of RNA transcription on the insertion element IS30 of E. coli K12. en. *EMBO J.* **4**, 2687–2693 (Oct. 1985).
60. Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J. & Feist, A. M. Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. en. *Cell Syst* **3**, 238–251.e12 (Sept. 2016).
61. Jensen, K. F. The Escherichia coli K-12 “wild types” W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. *J. Bacteriol.* **175**, 3401–3407 (1993).
62. Subbarayan, P. R. & Sarkar, M. A comparative study of variation in codon 33 of the rpoS gene in Escherichia coli K12 stocks: implications for the synthesis of σ s. *Mol. Genet. Genomics* **270**, 533–538 (Jan. 2004).
63. Vijayendran, C., Polen, T., Wendisch, V. F., Friehs, K., Niehaus, K. & Flaschel, E. The plasticity of global proteome and genome expression analyzed in closely related W3110 and MG1655 strains of a well-studied model organism, Escherichia coli-K12. *J. Biotechnol.* **128**, 747–761 (Mar. 2007).
64. Cho, B.-K., Kim, D., Knight, E. M., Zengler, K. & Palsson, B. O. Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states. en. *BMC Biol.* **12**, 4 (Jan. 2014).
65. Seo, S. W., Kim, D., Latif, H., O’Brien, E. J., Szubin, R. & Palsson, B. O. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in Escherichia coli. en. *Nat. Commun.* **5**, 4910 (Sept. 2014).
66. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. en. *Genome Biol.* **10**, R25 (Mar. 2009).

67. Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E. J., Zimmermann, M. T., Yan, H., Sun, Z., Zhang, Y., Wu, S. T., Huang, H., Wilson, M. D., Kocher, J.-P. A. & Li, W. MACE: model based analysis of ChIP-exo. en. *Nucleic Acids Res.* **42**, e156 (Nov. 2014).
68. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. en. *PLoS Comput. Biol.* **9**, e1003118 (Aug. 2013).
69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biol.* **15**, 550 (2014).
70. Kim, D. *Systems evaluation of regulatory components in bacterial transcription initiation* (University of California, San Diego, 2014).
71. Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I. & Karp, P. D. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. en. *Nucleic Acids Res.* **45**, D543–D550 (Jan. 2017).
72. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
73. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Kdd* **96** (1996), 226–231.
74. Gansner, E. R. & North, S. C. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exp.* **30**, 1203–1233 (2000).
75. Frigyesi, A., Veerla, S., Lindgren, D. & Höglund, M. Independent component analysis reveals new and biologically significant structures in micro array data. en. *BMC Bioinformatics* **7**, 290 (June 2006).
76. D’agostino, R. B., Belanger, A. & D’agostino, R. B. A Suggestion for Using Powerful and Informative Tests of Normality. *Am. Stat.* **44**, 316–321 (Nov. 1990).
77. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. en. *J. Neurosci. Methods* **134**, 9–21 (Mar. 2004).
78. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. en. *Bioinformatics* **19**, 185–193 (Jan. 2003).

79. Zou, H., Hastie, T. & Tibshirani, R. Sparse Principal Component Analysis. *J. Comput. Graph. Stat.* **15**, 265–286 (June 2006).

Chapter 3

Matrix factorization recovers consistent regulatory signals from disparate datasets

The availability of gene expression data has drastically increased in recent years. This data deluge could result in detailed inference of underlying regulatory networks, but the diversity of experimental platforms and protocols introduces critical biases that could hinder scalable analysis of existing data. Here, we show that the underlying structure of the *E. coli* transcriptome, as determined by Independent Component Analysis (ICA), is conserved across multiple independent datasets, including both RNA-seq and microarray datasets. We also show that echoes of this structure remain in the proteome, accelerating biological discovery through multi-omics analysis. We subsequently combined five transcriptomics datasets into a large compendium containing over 800 expression profiles, and discovered that its underlying ICA-based structure was

still comparable to that of the individual datasets. ICA thus enables deep analysis of disparate data to uncover new insights that were not visible in the individual datasets.

3.1 Background

Publicly available datasets, such as the NCBI Gene Expression Omnibus (GEO)[1] and Array Express [2], contain thousands of transcriptomics datasets that are often designed and analyzed for a specific study. Historically, microarrays were the platform of choice for transcriptomic interrogation. Over the past decade, usage of RNA sequencing (RNA-seq) has surpassed microarrays due to its higher sensitivity and ability to detect new transcripts [3].

Multiple consortia have performed extensive comparisons of expression levels across different microarray and RNA-seq platforms [4–6]. These studies showed that absolute gene expression levels cannot be accurately measured by either expression profiling technique; however, with appropriate quality controls in place, relative abundances are consistent across a wide range of transcriptomics platforms. Unfortunately, integration of omics datasets are often complicated by batch effects and technical heterogeneity, which are not easily corrected [7].

Differential expression analysis is the most common analytical method applied to transcriptomics datasets. However, differential expression analysis is limited in dimensionality, interpretability, and reproducibility; it can only be applied to pairs of experimental conditions, requires additional analysis to interpret large swaths of differentially expressed genes [8, 9], and is highly dependent on the quantification pipeline [10, 11]. Alternatively, machine learning methods, especially matrix factorization [12, 13], have provided new tools for extracting low-dimensional biological information from large omics data.

In particular, independent component analysis (ICA) has been shown to extract bio-

logically significant gene sets from many transcriptomics datasets [14–19] and two proteomics datasets [20]. ICA outperformed 42 module detection methods in a comprehensive examination across 5 organisms [21]. Previously, we applied ICA to a high-quality *E. coli* gene expression compendium to extract 92 independently-modulated groups of genes (called i-modulons). Of these 92 i-modulons, 61 exhibited high overlap with characterized regulons, and quantified the condition-dependent activities of the linked transcriptional regulators. An additional 25 i-modulons were linked to biological functions or genetic perturbations (leaving only 6 uncharacterized). These i-modulons have provided clear biological explanations for transcriptional changes in significantly perturbed cells [22, 23]. Other groups have successfully identified co-regulated gene sets using ICA on human transcriptomic datasets [16, 19], but many independent components could not be characterized, largely due to the high fraction of unknown human genes [24] compared to the model bacterium *E. coli* [25].

A recent study showed that among dimensionality reduction methods, ICA produced the most similar components across 14 independent cancer microarray datasets [26]. In this study we aim to determine if consistent regulatory components can be identified in expression datasets spanning disparate experimental conditions, and whether these components are robust to dataset integration. Through analysis of five independent *E. coli* gene expression datasets and two independent proteomics datasets, we identify a coherent structure without requiring any batch normalization procedures. In addition, integrated analysis of the different datasets demonstrated compelling evidence towards regulon discovery. These results present ICA as a promising tool to integrate and understand the flood of omics data challenging scientists today.

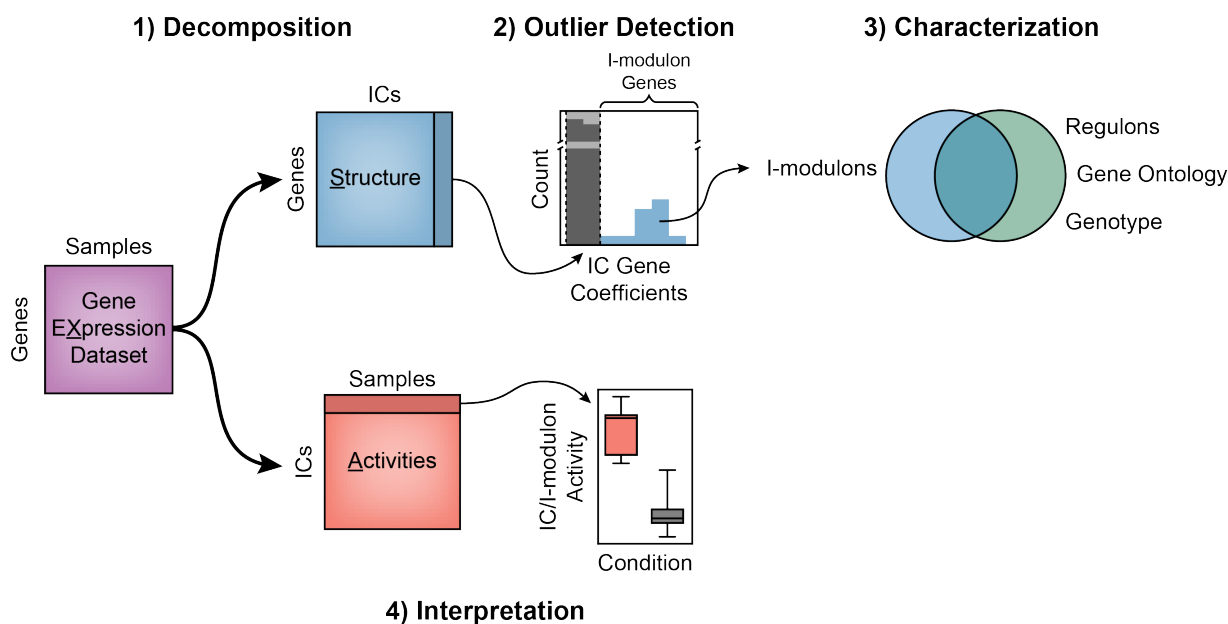
3.2 Results

3.2.1 ICA extracts i-modulons from five transcriptomic datasets

We compiled two RNA-seq and three microarray datasets, each using a different expression profiling technology or generated from a different research group (Table 3.1). Each dataset was independently processed, centered, and decomposed with ICA (See Methods). This process generated a set of independent components (ICs) for each dataset (Figure 3.1). Each IC contains a coefficient for every gene, although most gene coefficients were near zero for a given IC. To understand the biological role of each IC, we selected genes with sufficiently non-zero coefficients and refer to this set of genes as an i-modulon [27]. ICs, and their corresponding i-modulons, represent the underlying structure of the transcriptome under any experimental condition in the database. The condition-dependent dynamics of gene expression are captured by the activities of the ICs (also referred to as i-modulon activities). In this study, we focused on the condition-invariant structure of the transcriptome (i.e. ICs and their resultant i-modulons), to determine if the structure is conserved across different expression profiling platforms.

Table 3.1: Summary of transcriptomic datasets

Dataset	Platform(s)	Research Group	Samples	Unique Conditions	Total I-modulons
RNAseq-1 (PRECISE)	Illumina MiSeq, HiSeq, NextSeq & GAIIX	Bernhard Palsson	278	163	91
RNAseq-2	Applied Biosystems 5500XL Genetic Analyzer	Balász Papp & Csaba Pál	84	28	52
MA-1	Affymetrix <i>E. coli</i> Antisense Genome Array	Timothy Gardner	260	115	103
MA-2	Affymetrix <i>E. coli</i> Antisense Genome Array	Bernhard Palsson	124	39	58
MA-3	Affymetrix <i>E. coli</i> Genome 2.0 Array	Bernhard Palsson	56	20	32

**Figure 3.1:** Schematic illustration of the workflow applied to each data set.

We categorized the resulting i-modulons from each dataset into four classes, based on their gene content: (1) Regulatory, (2) Functional, (3) Genomic, and (4) Uncharacterized (Figure 3.2a,b). Prior work showed that i-modulons are highly consistent with known regulons, often containing genes known to be regulated by a single transcriptional regulator (e.g. transcription factor, sigma factor, or riboswitch). However, it has been observed that i-modulons may represent combinations of multiple regulons [16, 27].

In this study, a *Regulatory* i-modulon was defined as an i-modulon that was statistically enriched with a single regulon (Fisher’s exact test, $\text{FDR} < 10^{-5}$). Regulatory i-modulons were named after the single transcriptional regulator whose targets provided the best overlap (See Methods).

Functional i-modulons contained genes with highly similar functions, but lacked common regulators. For example, three i-modulons were identified in MA-1, MA-2, and RNAseq-1, respectively, that share 10 genes and were enriched in the gene ontology (GO) term “lipopolysaccharide core region biosynthetic process” (Figure 3.2c). These i-modulons were named based on the GO term with the lowest enrichment p-value ($\text{FDR} < .01$). This category also included i-modulons composed of prophages with no known regulators. Functional i-modulons represent a compelling opportunity for discovery of new transcriptional regulators [28].

Genomic i-modulons result from engineered overexpression or knock-out of one or more genes. The activities of these i-modulons represent the presence of these genomic alterations, such as large-scale duplications or deletions. These i-modulons can be useful to validate strain-specific mutations or deletions [27] and successful plasmid transformations [29]. Genomic i-modulons were categorized by reconciling i-modulon genes and activities with strain-specific genotypes.

The remaining *Uncharacterized* i-modulons could not be interpreted, either due to a high

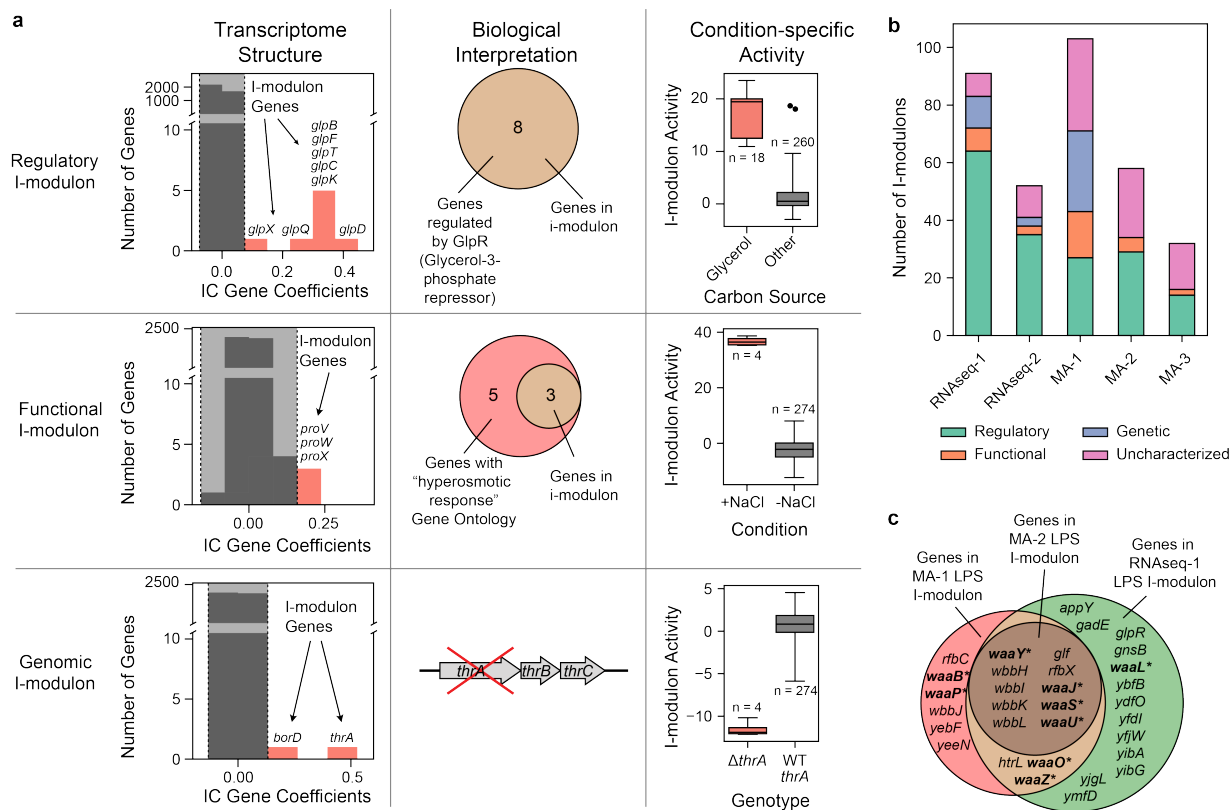


Figure 3.2: Characterization of i-modulons derived from five independent gene expression datasets. (a) Descriptions of the three classes of characterized i-modulons. The first column contains histograms illustrating the distribution of gene coefficients in each of three independent components (ICs) from the RNAseq-1 dataset. Genes outside of a threshold (in red) belong to an “i-modulon”. The second column illustrates the biological interpretation of the i-modulon types. I-modulons are characterized by comparing their genes with known regulons, ontological annotations, and genotypes. The third column displays the ICA-computed activity levels for the selected ICs across all 278 conditions in the RNAseq-1 dataset. (b) Bar chart describing the four types of i-modulons computed from each of the five gene expression datasets. (c) Comparison of three Functional i-modulons, each derived from decomposition of a different dataset. Asterisks indicate genes annotated with the GO term “lipopolysaccharide core region biosynthetic process”.

number of uncharacterized genes or the presence of seemingly unrelated genes. Uncharacterized i-modulons may represent undiscovered regulons, noise, or other unwanted sources of variation in the datasets.

The RNA-seq datasets produced the highest fraction of characterized i-modulons; only 9% of i-modulons in RNAseq-1, and 21% of i-modulons in RNAseq-2 were Uncharacterized. In

contrast, we were unable to characterize 31%, 41%, and 50% of the i-modulons derived from the microarray compendia (MA-1, MA-2, and MA-3, respectively). The lowest fraction of meaningful i-modulons was observed in MA-3, which also contained the fewest samples and unique conditions.

3.2.2 I-modulon structure is conserved across transcriptomic datasets

Each of the five transcriptomic datasets were created using different technologies or generated by different research groups, spanning fifteen years (2004-2019). In addition, each dataset contained vastly different experimental conditions and genotypes, such as overexpressed cellular division proteins in the MA-1 dataset, different nutritional supplements in the RNAseq-1 dataset, or evolved endpoints resistant to antibiotics in the RNAseq-2 dataset. Although the presence of different conditions in each dataset complicate the identification of consistent i-modulons, many of the i-modulons generated from the five datasets unexpectedly shared similar genes and annotations, as demonstrated by the lipopolysaccharide i-modulon discussed above (Figure 3.2c).

To assess the reproducible nature of i-modulons, we compared all i-modulons found in each dataset using a Reciprocal Best Hit (RBH) graph [26] (Figure 3.3a). In the RBH graph, each node represents an i-modulon, and nodes are connected when i-modulons from two different datasets find each other as the best scoring i-modulon in the other dataset [30]. I-modulons were scored by the absolute correlation between the gene coefficients in their respective ICs. For clarity, edges with low scores were removed from the graph (See methods, Figure C.1).

Of the 336 i-modulons identified across all five datasets, 151 i-modulons (45%) were linked to an i-modulon in another dataset, as shown in Figure 3.3a. Of these 151 reproducible i-modulons, 110 (73%) were classified as Regulatory (Figure 3.3b). Only one Genomic i-modulon was matched to an i-modulon from another dataset. This i-modulon captured the transcriptomic

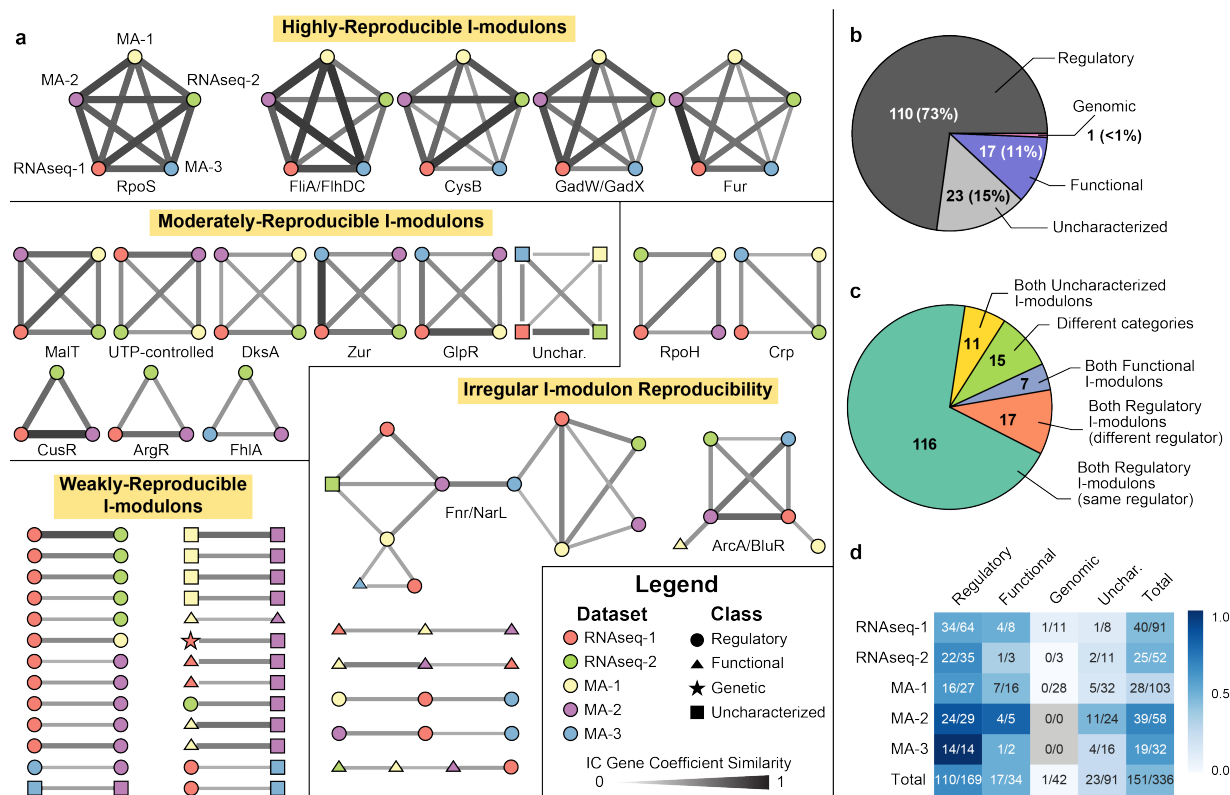


Figure 3.3: The i-modulon structure is conserved across five datasets. (a) Reciprocal best hit (RBH) graph indicating i-modulons as nodes and RBHs as edges. Node color indicates the source dataset for the i-modulon, as denoted for the RpoS i-modulon. Node shape indicates the i-modulon category. Edge thickness and darkness indicate the gene coefficient similarity. Clusters are labelled with the regulator(s) that are linked to the i-modulons in the cluster, if available, and grouped by level of i-modulon reproducibility. (b) Pie chart describing the categories of all i-modulons shown in the RBH graph. (c) Pie chart describing the types of edges in the RBH graph. (d) Heatmap indicating how many i-modulons from each dataset and category were in the RBH graph.

effects of a genomic duplication that occurred during an adaptive laboratory evolution [27, 31].

This Genomic i-modulon was linked to an i-modulon in the MA-2 dataset that was activated for strains evolved under a different selection pressure [32] (Figure C.2a,b). Although the strains in the MA-2 dataset were not resequenced, the i-modulon gene coefficients indicate that these strains likely acquired an identical duplication along their evolutionary trajectory (Figure C.2c).

This example showed how analysis of historical data with new context can create new knowledge.

Of the 166 connections found between i-modulons, 133 (80%) were between two Regulatory i-modulons, and 116 of these 133 edges (87%) connected i-modulons enriched in the same regulator (Figure 3.3c). Across all five datasets, Regulatory i-modulons were the most reproducible category (Figure 3.3d). The lowest fraction of reproducible Regulatory i-modulons was observed in the RNAseq-1 dataset, likely due to the activation of niche transcription factors that are only active under specific environmental conditions.

Many i-modulons in the RBH graph separated into well-defined clusters, each linking at most one i-modulon from each dataset. I-modulons in a cluster were often linked to the same regulator, as labelled below each cluster. However, some clusters exhibited irregular connectivity, the largest of which contained i-modulons linked to either of the nitrate response regulators Fnr and NarL. Cellular respiration is regulated by a complex combination of highly interconnected transcription factors [33], indicating that these i-modulons may represent combined effects of regulators. Each dataset contains a different set of conditions, which can activate different groups of respiration-related genes leading to different co-expression patterns between datasets (Figure C.3).

The RBH graph contained five highly-connected clusters, each containing one i-modulon from each of the five datasets. Three of these clusters were enriched with genes from a single regulon: CysB, RpoS, and Fur. The other two clusters contained i-modulons enriched with genes regulated by closely-related transcriptional regulators. The i-modulons in one cluster were enriched with genes regulated by the motility-related regulators FliA and FlhDC, and the i-modulons in other cluster were enriched with genes regulated by the acid stress response transcription factors GadW and GadX.

The IC gene coefficients for the i-modulons within a cluster were highly correlated, indi-

cating that the genes in the i-modulons were modulated at similar ratios across all five datasets, regardless of expression profiling platform. This correlation was demonstrated by the i-modulons in the CysB i-modulon cluster (Figure 3.4a). CysB is a transcription factor that regulates cysteine biosynthesis [34]. The IC gene coefficient similarities between i-modulons in the CysB cluster were above 0.5, except for comparisons involving the CysB i-modulon from the MA-3 dataset. On average, this CysB i-modulon captured 79% of the genes in the CysB i-modulons from other datasets, but contained an additional 71 genes that were absent from any other CysB i-modulon. Many of these genes were involved in the biosynthesis of other amino acids, and were present in different i-modulons in the other datasets. In fact, the MA-3 CysB i-modulon can be approximated ($R^2 = 0.47$) by a linear combination of 10 i-modulons from the RNAseq-1 dataset (Figure 3.4b).

3.2.3 ICA finds similar structures in the proteome

Since the i-modulon structure was highly consistent across transcriptomics datasets, we asked whether shadows of the transcriptional regulatory program could be observed in the proteome. Thus, we performed the extended ICA algorithm on two high-quality proteomics datasets from two independent research groups [35, 36]. The Proteomics-1 dataset had higher gene coverage (Figure C.4a) across 22 experimental conditions whereas the Proteomics-2 dataset included more conditions.

The Proteomics-1 dataset produced 19 i-modulons, whereas the Proteomics-2 dataset produced 22 i-modulons (Figure 3.5a). Ten of 19 i-modulons from the Proteomics-1 dataset and eight of 22 i-modulons from the Proteomics-2 dataset were linked to an i-modulon from the RNAseq-1 dataset using the RBH method, and two pairs of i-modulons were linked between the

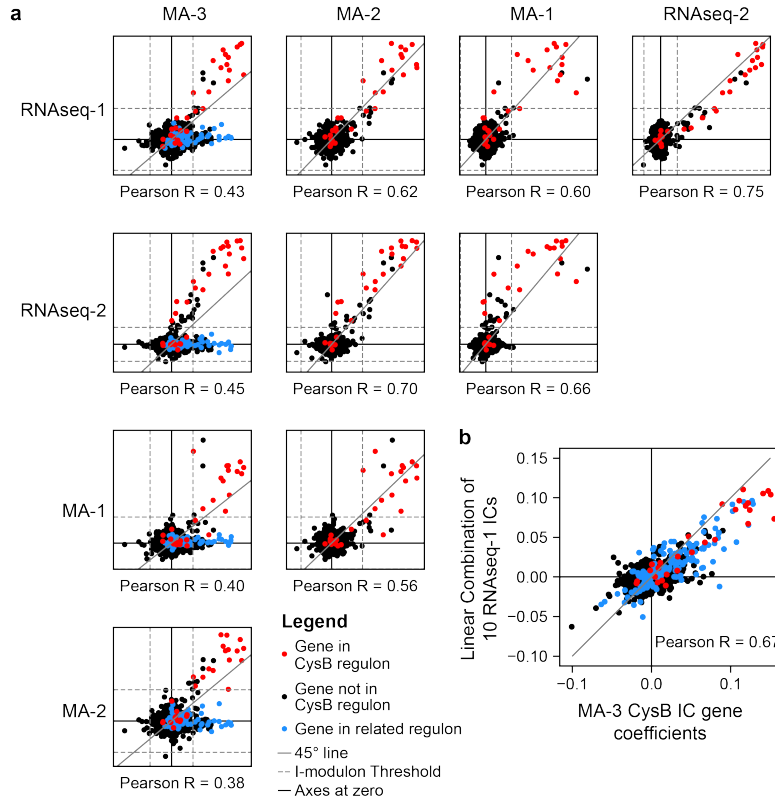


Figure 3.4: Comparison of IC gene coefficients for i-modulons enriched with genes in the CysB regulon. (a) Scatter plot between IC gene coefficients for CysB-linked i-modulons across all five datasets. Genes in the published CysB regulon are colored in red. For comparisons in the first column involving the MA-3 dataset, genes regulated by any of the following regulators are colored in blue: MetJ, TrpR, GlpR, ArgR, Lrp, CysB, leu-tRNA-mediated transcriptional attenuation, or thiamine riboswitch. All other genes are black. Dashed lines indicate i-modulon thresholds. Gray solid lines indicate the 45 degree line of equal gene coefficients. (b) Scatter plot between IC gene coefficients for the CysB i-modulon in MA-3 compared to a linear combination of 10 ICs from RNAseq-1. Color scheme is identical to panel (a).

proteomics datasets (Figure 3.5b). We note that most i-modulons from the proteomics datasets were Uncharacterized, likely due to the intrinsically lower coverage and smaller sizes of the datasets.

One pair of Uncharacterized i-modulons were independently detected in the RNAseq-1 dataset and the Proteomics-1 dataset (Figure C.4b). Due to the lower gene coverage in the Proteomics-1 dataset, we investigated the genes in the relevant RNAseq-1 i-modulon (Fig-

ure 3.5c), which were mostly localized in the periplasm and inner membrane. To gain a better understanding of these i-modulons, we inspected the i-modulon activities in both datasets and found high activity under osmotic stress and in laboratory evolution strains (Figure 3.5d,e). In addition, many genes in this i-modulon had been identified in two prior studies as genes up-regulated by RcsC [37], or upregulated in response to beta-lactam antibiotics [38]. RcsC is a histidine kinase in the Rcs phosphorelay system that responds to envelope stress [38]. Therefore, we propose that all the genes in this i-modulon are co-regulated by a transcription factor related to the Rcs system that may currently be uncharacterized. This represents a clear potential for discovery.

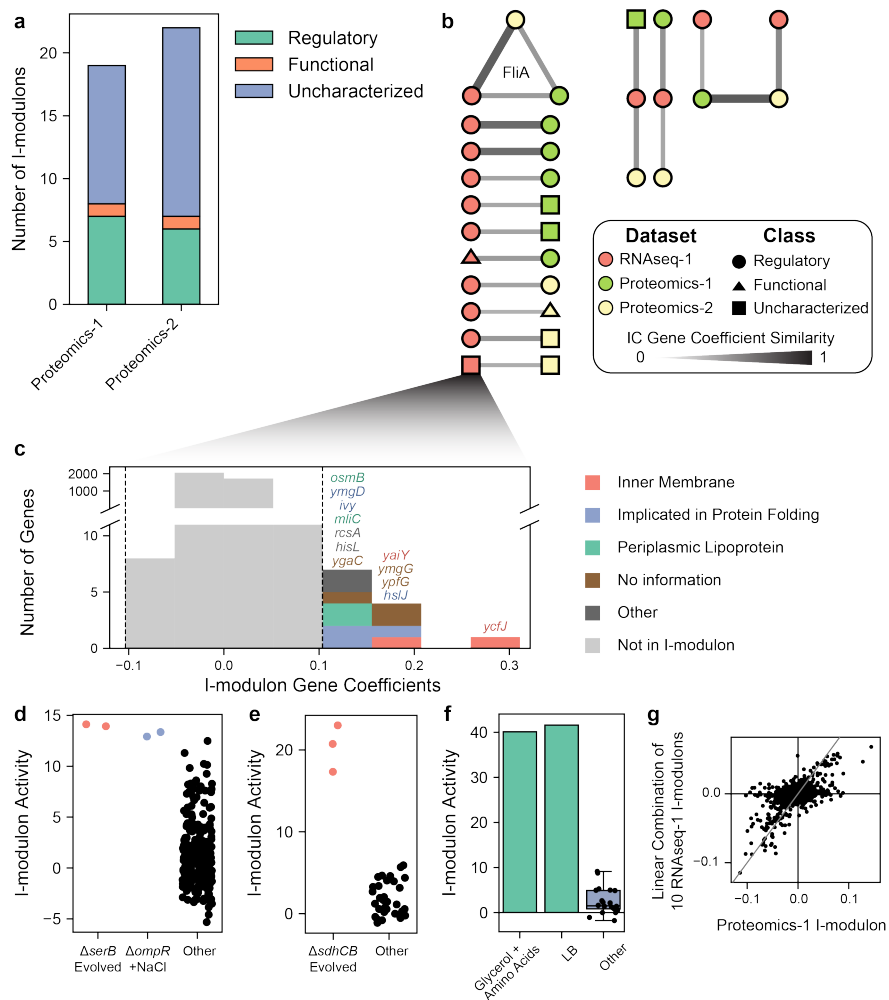


Figure 3.5: ICA elucidates similar structures in the *E. coli* transcriptome and proteome. (a) Bar chart showing the types of i-modulons extracted from two independent proteomic datasets. (b) RBH graph between i-modulons from two proteomics datasets and the RNAseq-1 i-modulons. (c) Genes in the RNAseq-1 i-modulon that was also detected in the Proteomics-1 dataset (d) Activity of the Uncharacterized i-modulon illustrated in panel (c). (e) Activity of the Proteomics-2 i-modulon linked to the RNAseq-1 i-modulon illustrated in panel (c) (f) Activity levels of an Uncharacterized i-modulon in the Proteomics-1 dataset. (g) Scatter plot between IC protein coefficients for the i-modulon described in panel (f) compared to a linear combination of 10 ICs from RNAseq-1. Black lines indicate zero IC gene coefficients, and gray line indicates 45 degree line.

It is important to note that the IC gene coefficient similarities in the RBH graph between the proteomics and the RNAseq-1 dataset were lower than the gene coefficient similarities in the transcriptomic RBH graph (Mann-Whitney U Test, p -value = 0.005). Although some of this deviation may be due to differences between relative abundances between transcripts and proteins, the small sample sizes of the proteomics datasets resulted in i-modulon coalescence as previously described with the MA-3 CysB i-modulon. In particular, one Uncharacterized i-modulon from the Proteomics-1 dataset was only active under two media conditions: LB rich media, and Glycerol M9 media supplemented with amino acids (Figure 3.5f). This proteomic i-modulon can be loosely approximated ($R^2 = 0.33$) by a linear combination of 10 RNAseq-1 i-modulons, (Figure 3.5g) related to amino acid biosynthesis and catabolite repression. Six of these i-modulons coincide with the 10 RNAseq-1 i-modulons comprising the MA-3 CysB i-modulon.

Due to the lower coverage of genes and low number of conditions in the proteomics datasets, it is not currently possible to conclusively state that the i-modulon structure is conserved between proteomics and transcriptomics datasets, but this section provides evidence that a similar structure may exist.

3.2.4 Data integration reveals increased resolution and novel insights

I-modulons are a linear approximation of the structure underlying each omics dataset analyzed in this study. Since elements of this structure were highly consistent across the five transcriptomics datasets, we combined the datasets together to explore whether ICA would identify a similar structure in the combined compendium, which contained 802 expression profiles (Figure 3.6a).

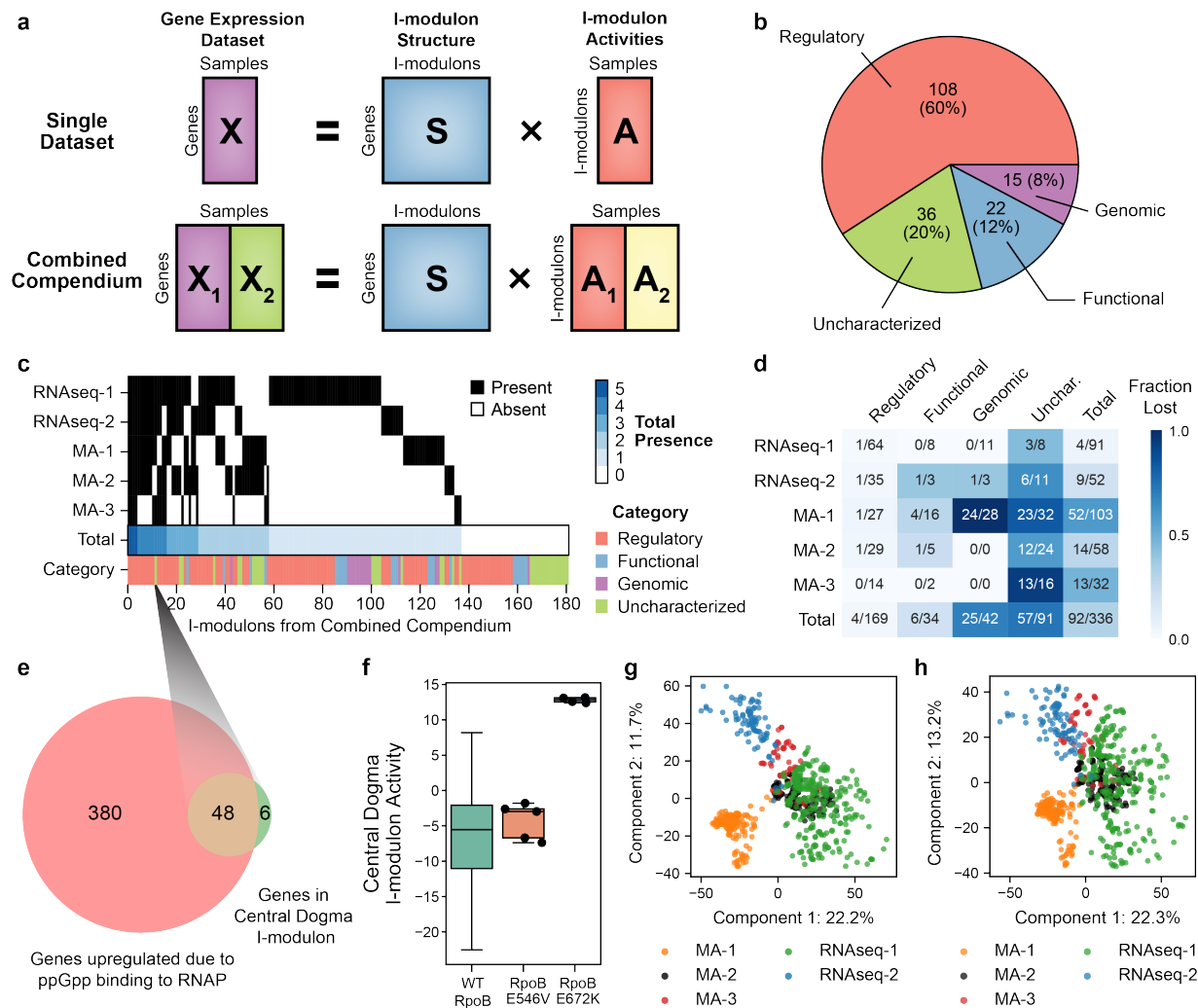


Figure 3.6: Decomposition of combined compendium results in increased resolution. (a) Schematic illustration of data integration. (b) Pie chart showing the categories of the 181 i-modulons from the combined dataset. (c) Heatmap illustrating which i-modulons from the combined compendium were matched to i-modulons in individual datasets using the RBH method. The total number of matches for each i-modulon is shown in blue. Category of each i-modulon is shown below. (d) Heatmap illustrating how many i-modulons from each category and dataset were matched to an i-modulon in the full dataset using the RBH method. (e) Overlap between genes in an Uncharacterized i-modulon from the combined compendium and genes upregulated by ppGpp binding to RNA Polymerase. (f) Boxplot of the Central Dogma i-modulon activities for expression profiles in the RNAseq-1 dataset. (g) Principal component loadings of the combined compendium expression levels without batch correction (h) Principal component loadings of the i-modulon activities from the combined compendium.

We extracted 181 i-modulons from the combined compendium, and characterized them as described previously (Figure 3.6b). The majority of i-modulons from this combined compendium (60%) were enriched with targets of a known transcriptional regulator. To understand the effects of dataset integration, we compared the ICs obtained from the combined compendium to the ICs found in the original independent datasets using the RBH method.

First, we asked whether the i-modulons identified in the original datasets were retained upon data integration. From the RBH graph, 75% of the 181 i-modulons in the combined compendium could be directly linked back to at least one i-modulon derived from an individual dataset, many of which were linked to i-modulons from two or more datasets (Figure 3.6c). The two datasets with the largest number of unique conditions, (RNAseq-1 and MA-1) contained the most uniquely mapped i-modulons, as these datasets were more likely to activate niche transcriptional regulators.

We next investigated the properties of i-modulons that were not retained upon data integration. Nearly all of the 92 i-modulons that did not appear in the RBH graph were categorized as either Genomic or Uncharacterized (Figure 3.6d). Half of the lost i-modulons were Genomic i-modulons from the MA-1 dataset, which comprised mainly of expression profiles from single gene knock-outs. The i-modulons that were not retained upon data integration each account for a significantly lower fraction of expression variance than an i-modulon that was retained (Mann-Whitney-U Test p -value $< 10^{-5}$, Figure C.5a). This analysis indicated that the missing i-modulons represent weak signals, or potentially expression noise, that lose their ability to be detected in the presence of stronger signals.

Additionally, 44 i-modulons in the combined compendium could not be traced back to the i-modulons from the individual datasets. Of these new 44 i-modulons, 21 represented the

effects of transcriptional regulators that could not be independently discriminated in any of the individual datasets alone (Figure C.5b), and 12 were dominated by a single gene (Figure C.5c). A previous study showed that over-decomposition of a transcriptomic dataset results in smaller groups of genes found in each i-modulon, but does not affect the biological relevance of other i-modulons [39].

One Uncharacterized i-modulon in the combined compendium was linked back to four of the five individual datasets. These four i-modulons created the only uncharacterized cluster in Figure 3.3a. Although this i-modulon was not enriched in any GO term, it contained many genes encoding functions related to the central dogma of molecular biology, such as rRNA and tRNA modification, RNases, and helicases (Figure C.5d). We henceforth refer to this i-modulon as the “Central Dogma” i-modulon. 63% of genes in this i-modulon are not known to be regulated by any transcription factor, and the remaining genes are not enriched in any regulator either.

A recent study found that binding of ppGpp, the stringent response alarmone [40], to RNAP directly upregulated 428 genes [41], including 48 of the 54 genes in Central Dogma i-modulon (Figure 3.6e). Previously, we have investigated the effects of a point mutation in RpoB, the β subunit of RNAP, which results in increased growth and decreased expression of stress-response genes [27, 42]. Expression profiles for the RpoB E672K mutant strain produce a distinctly higher activity for the Central Dogma i-modulon than any other condition or strain in the RNAseq-1 dataset (Figure 3.6f), providing evidence that the RpoB E672K mutation affects ppGpp binding to RNAP, even though the mutation occurs over 25 Å from the binding site.

Next, we investigated whether dataset integration could change the quality of the Regulatory i-modulons. Since a Regulatory i-modulon consists of genes in a known regulon, the quality of a Regulatory i-modulon can be assessed using the F1-score, which is the harmonic mean of

precision and recall of the i-modulon compared to the regulon (Figure C.6a). We inspected all Regulatory i-modulons that were found in both an individual dataset and the combined compendium, and found that the average F1-score increased from 0.54 to 0.59 (Wilcoxon Rank Sum p -value = 10^{-5}) (Figure C.6b). This result showed that on average, Regulatory i-modulons from the combined compendium were more similar to the previously known regulons than the Regulatory i-modulons from individual datasets.

Finally, we checked whether the i-modulon activities were affected by dataset integration. When we apply ICA to an expression profiling dataset, we obtain the **S** matrix, encoding the i-modulon structure, and an **A** matrix, which contains activity levels for each i-modulon across every expression profile (Figure 3.1 and Figure 3.6a). I-modulon activities reflect the overall change in expression of the i-modulon genes, and serve as a proxy for transcription factor activities [23]. Therefore, it is important to ensure that the relative i-modulon activities are unchanged upon dataset integration. The absolute Pearson R correlation coefficient between the i-modulon activities of linked components showed that most i-modulon activities are mostly unchanged upon dataset integration (Figure C.6c). However, principal component analysis of the full expression compendium (Figure 3.6g) and i-modulon activities (Figure 3.6h) illustrated a clear batch effect between platforms. Therefore i-modulon activities cannot be compared across datasets, but can still reliably be compared within the same dataset.

3.3 Discussion

Matrix decomposition is a powerful approach to extract knowledge from large transcriptomics datasets. In particular, we have shown that ICA identifies highly similar structures between dissimilar datasets for the model bacteria *E. coli*. In addition, a combined compendium

produced many identical i-modulons as the individual datasets, and could distinguish further signals that could not be identified in the separate datasets. The i-modulons derived from the compendium can be applied to interpret new datasets, accelerating discovery and providing a standard framework that could be used to investigate any transcriptional regulator. Throughout this study, we observed various properties of the i-modulon decompositions: (1) i-modulons co-occurring in multiple datasets tend to represent the effects of transcriptional regulators, (2) i-modulon detection from a dataset is dependent on the experimental conditions used to generate it, (3) transcriptomic and proteomic i-modulons are surprisingly similar, (4) ICA can be applied to cross-platform transcriptomic compendia without the need for normalization, and (5) integration of data sets reveals i-modulons not found in individual data sets.

In all five transcriptomic *E. coli* datasets, most i-modulons could be characterized as representing the effects of a transcriptional regulator or a gene knock-out (i.e. Regulatory or Genomic). However, when extending to less-characterized organisms, it could be difficult to determine whether the remaining i-modulons are technical artifacts or contain true biological insight, as demonstrated by the Central Dogma i-modulon. This example demonstrates that if an i-modulon is identified in multiple datasets, or if an i-modulon persists upon addition of new datasets, then it could represent a true transcriptional signal.

Another important note is that the i-modulons extracted from each dataset are sensitive to the experimental conditions that are represented in the dataset; ICA cannot extract an i-modulon for a transcription factor whose activity never changes. Additionally, the CysB i-modulons demonstrated that i-modulons may represent the effects of multiple regulators, when the activities of the regulators are highly correlated across the measured conditions. However, adding new data or increasing the dimensionality of the decomposition can decouple the regulators, splitting the

i-modulon into its biological parts [27, 39].

Therefore, an important design decision for omics compendia is to ensure enough diverse conditions such that each transcriptional regulator has its own i-modulon. However, once a compendium reaches a critical size, over-decomposition of the compendium may result in smaller, less-relevant i-modulons, such as i-modulons that track the expression of a single gene. The optimal dimensionality of the transcriptome is an open question that has been investigated [39, 43], but warrants further attention in the context of prokaryotic transcriptomes and proteomes.

In contrast to the condition-invariant i-modulon structure, i-modulon activities represent the condition-dependent dynamics of expression profiles. In this study, we do not apply any normalization techniques to the data and clearly observe batch effects in the activity matrix for the combined compendium. Further work comparing identical experimental conditions from separate protocols is required to enable i-modulon activity comparisons across datasets.

We have shown that the *E. coli* transcriptome contains a conserved, underlying structure that is found across multiple independent datasets and extends into the proteome. This powerful observation enables unprecedented re-analyses of thousands of previously published datasets and demonstrates how data science can unlock hidden potential in complex biological datasets.

3.4 Methods

3.4.1 RNA-seq and microarray data processing

The full, log-transformed transcripts-per-million (log-TPM) for the RNAseq-1 dataset [27] was obtained from <https://github.com/SBRG/precise-db>. Raw data comprising the RNAseq-2 compendium were obtained from NCBI SRA under the Bioproject accession number PR-

JNA379428 [44]. Raw data was processed using as described in Sastry et al. [27].

CEL files were obtained for the three microarray datasets from NCBI GEO [1]. In order to build large enough datasets for analysis, datasets MA-2 and MA-3 included all public data available from our research group that used the same expression platform. Each microarray dataset was normalized using robust multichip average (RMA) with default arguments from the R package *affy*[45].

As the probes in microarrays may vary across platforms, only genes that were measured in all 5 datasets (3880 genes) were included in the analysis. Low quality expression profiles that clustered separately from the rest of the dataset were removed from the MA-1 and MA-2 datasets. Each dataset was individually centered by subtracting the average expression profile across the replicates of a dataset-specific reference condition. To create the combined compendium, the centered datasets were concatenated, without any additional normalization.

Only genes measured in all five datasets were retained, resulting in 3880 genes per dataset. Gene names, b-numbers, operons, and descriptions were obtained from Ecocyc [46]. Clusters of orthologous group (COG) annotations were obtained from eggNOG 4.5.1 [47]. Additional annotations were obtained from Gene Ontology [48]. The transcriptional regulatory network (TRN) for transcription factors, small RNAs, and sigma factors was obtained from RegulonDB v10.0 [49], and supplemented with newly identified regulons from recent publications [27, 28]. Riboswitches, tRNA-mediated attenuation, and dksA binding sites were obtained from Ecocyc [46], and UTP/CTP-dependent attenuation and reiterative transcription were obtained from Turnbough and Switzer [50].

3.4.2 Proteomics data processing

Processed data for the Proteomics-1 dataset was acquired from Schmidt et al. [35], using the values reported for protein copies per cell. Only genes with measurements across all samples were retained, resulting in 2058 genes. Processed data for the Proteomics-2 dataset was acquired from Heckmann et al. [36], and augmented with additional samples for growth on different carbon sources that were acquired as described in Heckmann et al. [36]. Protein abundance estimation is described in Heckmann et al. [36]; in short, the top3 metric[51, 52] was calibrated using the UPS2 standard to obtain protein amount loaded based on the average of three technical replicates per sample. For each strain, two biological replicates were profiled and abundances were averaged before conducting ICA.

3.4.3 Independent Component Analysis

ICA decomposes a matrix ($\mathbf{X}$) into two matrices: $\mathbf{S}$ contains the independent signals, or structure, of the dataset, and $\mathbf{A}$ contains the condition-dependent activities of the signals:

$$\mathbf{X} = \mathbf{SA}$$

ICA was performed on each individual dataset and the combined compendium, as described in Sastry et al.[27]. Briefly, we executed FastICA 100 times with random seeds and a convergence tolerance of 10^{-6} for RNA-seq data, and a convergence tolerance of 10^{-7} for proteomics data. We constrained the number of independent components (ICs) in each iteration to the number of components that reconstruct 99% of the variance as calculated by principal component analysis. The resulting ICs were clustered using DBSCAN to identify robust ICs, with parameters with epsilon of 0.1, and minimum cluster seed size of 50. This process was

repeated 10 times, and only ICs that consistently occurred in all runs were kept.

As described in Sastry et al.[27], i-modulons were extracted from ICs by iteratively removing genes with the largest absolute value and computing the D’Agostino K^2 test statistic of the resulting distribution. Once the test statistic fell below a cutoff, we designated the removed genes as the “i-modulon”.

To identify this cutoff for each individual dataset, we performed a sensitivity analysis on the concordance between significant genes in each IC and all known regulons. First, we isolated the 20 genes from each IC with the highest absolute gene coefficients. We then compared each gene set against all known regulons using the one-sided Fisher’s Exact Test ($\text{FDR} < 10^{-5}$). For each component with at least one significant enrichment, we selected the regulator with the lowest p-value.

Next, we varied the D’Agostino K^2 test statistic from 200 through 1000 in increments of 50. Using the protocol defined above, i-modulons were extracted from ICs at each test statistic value, and the F1-score (harmonic average between precision and recall) was computed between the significant genes and its linked regulator. The test statistic with the maximum F1-score was used as the test statistic cutoff for the respective dataset.

The above methods were also used for the proteomics datasets. The TRN and GO terms were trimmed to only include proteomics genes, and the FDR for Fisher’s Exact Test was lowered to 10^{-3} to account for the lower protein coverage.

3.4.4 Characterizing i-modulons

We compared the set of significant genes in each i-modulon to each regulon (defined as the set of genes regulated by any given regulator) using the one-sided Fisher’s Exact Test (FDR

$< 10^{-5}$ for transcriptomic i-modulons, and $\text{FDR} < 10^{-3}$ for proteomic i-modulons). We then compared the significant genes in each i-modulon to the genes in each Gene Ontology (GO) term using the one-sided Fisher’s Exact Test ($\text{FDR} < .01$). We added prophage information from Ecocyc to our GO database to capture i-modulons representing prophages. In general, the final annotation was selected by the enrichment with the lowest q-value. Some i-modulon annotations were manually curated. Genomic i-modulons were also manually curated by (1) comparing i-modulon genes to known genetic alterations (e.g. knock-outs or overexpression), and (2) validating that the i-modulon activities were affected in the appropriate direction for the corresponding strain.

3.4.5 Comparing i-modulon structures

To compare the complete structure of the transcriptomic datasets, we constructed the reciprocal best hit (RBH) graphs using the full IC gene coefficients, rather than just the i-modulon genes. We generated the RBH graph as described in Cantini et al. 2019 [26], using the following distance metric to compare ICs:

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

where $\rho_{x,y}$ is the Pearson correlation between components x and y .

We pruned all RBHs to remove links between ICs with similarities below 0.3. The full graph is shown in Figure C.1. The RBH graph was plotted using GraphViz [53].

3.4.6 Linear regression of i-modulons

Regression of the MA-3 CysB i-modulon and the Proteomics-1 i-modulon were performed using the LinearRegression function from Scikit-learn[54]. The ten ICs from the RNAseq-1

dataset with the highest absolute IC gene correlations with each of the CysB i-modulon and Proteomics-1 i-modulon were selected as regression variables for their respective regressions.

Acknowledgements

The authors would like to thank Dr. Dan Zielinski, Prof. Eivind Almaas and Yara Seif, and CJ Norsigian for informative discussions. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was funded by the Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (grant number NNF10CC1016517).

Author Contributions: A.V.S. and B.O.P. designed the study. A.V.S., A.H., and D.H. performed the analysis. E.K. and S.P. contributed to the interpretation of the data. A.V.S., D.H., E.K., S.P., and B.O.P. wrote and edited the manuscript.

Chapter 3, in part, is currently being prepared for submission for publication: **AV Sastry**, A Hu, D Heckmann, S Poudel, E Kavvas, and BO Palsson. 2019. “Matrix factorization recovers consistent regulatory signals from disparate datasets”. The dissertation author is the primary author.

3.5 References

1. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (Jan. 2013).
2. Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Papatheodorou, I., Sarkans, U. & Brazma, A. ArrayExpress update - from bulk to single-cell expression data. en. *Nucleic Acids Res.* **47**, D711–D715 (Jan. 2019).

3. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. en. *PLoS One* **9**, e78644 (Jan. 2014).
4. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. en. *Nat. Biotechnol.* **24**, 1151–1161 (Sept. 2006).
5. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. en. *Nat. Biotechnol.* **32**, 903–914 (Sept. 2014).
6. 't Hoen, P. A. C., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., Laros, J. F. J., Buermans, H. P. J., Karlberg, O., Brännvall, M., GEUVADIS Consortium, den Dunnen, J. T., van Ommen, G.-J. B., Gut, I. G., Guigó, R., Estivill, X., Syvänen, A.-C., Dermitzakis, E. T. & Lappalainen, T. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. en. *Nat. Biotechnol.* **31**, 1015–1022 (Nov. 2013).
7. Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology* **35**. Special Issue: Computation and Modeling, 498–507. ISSN: 0167-7799. <http://www.sciencedirect.com/science/article/pii/S0167779917300367> (2017).
8. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. en. *Nucleic Acids Res.* **37**, 1–13 (Jan. 2009).
9. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W. & Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. en. *Nucleic Acids Res.* **44**, W90–7 (July 2016).
10. Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Li, S., Mason, C. E., Olson, S., Pervouchine, D., Sloan, C. A., Wei, X., Zhan, L. & Irizarry, R. A. A benchmark for RNA-seq quantification pipelines. en. *Genome Biol.* **17**, 74 (Apr. 2016).
11. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. en. *BMC Bioinformatics* **14**, 91 (Mar. 2013).
12. Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., Goff, L. A., Li, Y., Ngom, A., Ochs, M. F., Xu, Y. & Fertig, E. J. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (Oct. 2018).
13. Way, G. P. & Greene, C. S. Discovering Pathway and Cell Type Signatures in Transcriptomic Compendia with Machine Learning. *Annu. Rev. Biomed. Data Sci.* **2**, 1–17 (July 2019).
14. Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U. & Zinovyev, A. Independent

- Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. en. *Int. J. Mol. Sci.* **20** (Sept. 2019).
15. Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T. & Huang, X. A review of independent component analysis application to microarray gene expression data. en. *Biotechniques* **45**, 501–520 (Nov. 2008).
 16. Karczewski, K. J., Snyder, M., Altman, R. B. & Tatonetti, N. P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. en. *PLoS Genet.* **10**, e1004122 (Feb. 2014).
 17. Engreitz, J. M., Daigle Jr, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. en. *J. Biomed. Inform.* **43**, 932–944 (Dec. 2010).
 18. Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R. & Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. en. *PLoS Comput. Biol.* **3**, e161 (Aug. 2007).
 19. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A. & Radvanyi, F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. en. *Cell Rep.* **9**, 1235–1245 (Nov. 2014).
 20. Liu, W., Payne, S. H., Ma, S. & Fenyő, D. Extracting Pathway-level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis. en. *Mol. Cell. Proteomics* **18**, S169–S182 (Aug. 2019).
 21. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. en. *Nat. Commun.* **9**, 1090 (Mar. 2018).
 22. Anand, A., Chen, K., Yang, L., Sastry, A. V., Olson, C. A., Poudel, S., Seif, Y., Hefner, Y., Phaneuf, P. V., Xu, S., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. en. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 25287–25292 (Dec. 2019).
 23. Anand, A., Chen, K., Catoiu, E., Sastry, A. V., Olson, C. A., Sandberg, T. E., Seif, Y., Xu, S., Szubin, R., Yang, L., Feist, A. M. & Palsson, B. O. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. en. *Mol. Biol. Evol.* (Oct. 2019).
 24. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. en. *PLoS Biol.* **16**, e2006643 (Sept. 2018).

25. Ghatak, S., King, Z. A., Sastry, A. & Palsson, B. O. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. en. *Nucleic Acids Res.* **47**, 2446–2454 (Mar. 2019).
26. Cantini, L., Kairov, U., de Reyniès, A., Barillot, E., Radvanyi, F. & Zinovyev, A. Assessing reproducibility of matrix factorization methods in independent transcriptomes. en. *Bioinformatics* (Apr. 2019).
27. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. The Escherichia coli transcriptome mostly consists of independently regulated modules. en. *Nat. Commun.* **10**, 5536 (Dec. 2019).
28. Gao, Y., Yurkovich, J. T., Seo, S. W., Kabimoldayev, I., Dräger, A., Chen, K., Sastry, A. V., Fang, X., Mih, N., Yang, L., Eichner, J., Cho, B.-K., Kim, D. & Palsson, B. O. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. en. *Nucleic Acids Res.* (Aug. 2018).
29. Tan, W. K. J. *Multi-omic characterization of E. coli for the purpose of microbial-based production* PhD thesis (UC San Diego, 2019).
30. Ward, N. & Moreno-Hagelsieb, G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? en. *PLoS One* **9**, e101850 (July 2014).
31. Guzmán, G. I., Sandberg, T. E., LaCroix, R. A., Nyerges, Á., Papp, H., de Raad, M., King, Z. A., Hefner, Y., Northen, T. R., Notebaart, R. A., Pál, C., Palsson, B. O., Papp, B. & Feist, A. M. Enzyme promiscuity shapes adaptation to novel growth substrates. en. *Mol. Syst. Biol.* **15**, e8462 (Apr. 2019).
32. Fong, S. S., Joyce, A. R. & Palsson, B. Ø. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. en. *Genome Res.* **15**, 1365–1372 (Oct. 2005).
33. Federowicz, S., Kim, D., Ebrahim, A., Lerman, J., Nagarajan, H., Cho, B.-K., Zengler, K. & Palsson, B. Determining the control circuitry of redox metabolism at the genome-scale. en. *PLoS Genet.* **10**, e1004264 (Apr. 2014).
34. Kredich, N. M. The molecular basis for positive regulation of cys promoters in Salmonella typhimurium and Escherichia coli. en. *Mol. Microbiol.* **6**, 2747–2753 (Oct. 1992).
35. Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R. & Heinemann, M. The quantitative and condition-dependent Escherichia coli proteome. en. *Nat. Biotechnol.* **34**, 104–110 (Jan. 2016).
36. Heckmann, D., Campeau, A., Lloyd, C. J., Phaneuf, P., Hefner, Y., Carrillo-Terrazas, M., Feist, A. M., Gonzalez, D. J. & Palsson, B. O. *Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers* en. Sept. 2019.

37. Ferrières, L. & Clarke, D. J. The RcsC sensor kinase is required for normal biofilm formation in *Escherichia coli* K-12 and controls the expression of a regulon in response to growth on a solid surface. en. *Mol. Microbiol.* **50**, 1665–1682 (Dec. 2003).
38. Laubacher, M. E. & Ades, S. E. The Rcs phosphorelay is a cell envelope stress response activated by peptidoglycan stress and contributes to intrinsic antibiotic resistance. en. *J. Bacteriol.* **190**, 2065–2074 (Mar. 2008).
39. Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E. & Zinovyev, A. Determining the optimal number of independent components for reproducible transcriptomic data analysis. en. *BMC Genomics* **18**, 712 (Sept. 2017).
40. Srivatsan, A. & Wang, J. D. Control of bacterial transcription, translation and replication by (p)ppGpp. en. *Curr. Opin. Microbiol.* **11**, 100–105 (Apr. 2008).
41. Sanchez-Vazquez, P., Dewey, C. N., Kitten, N., Ross, W. & Gourse, R. L. Genome-wide effects on *Escherichia coli* transcription from ppGpp binding to its two sites on RNA polymerase. en. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8310–8319 (Apr. 2019).
42. Utrilla, J., O’Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. en. *Cell Syst* **2**, 260–271 (Apr. 2016).
43. Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. *Sequential compression of gene expression across dimensionalities and methods reveals no single best method or dimensionality* en. Sept. 2019.
44. Lázár, V., Martins, A., Spohn, R., Daruka, L., Grézal, G., Fekete, G., Számel, M., Jangir, P. K., Kintsés, B., Csörgő, B., Nyerges, Á., Györkei, Á., Kincses, A., Dér, A., Walter, F. R., Deli, M. A., Urbán, E., Hegedűs, Z., Olajos, G., Méhi, O., Bálint, B., Nagy, I., Martinek, T. A., Papp, B. & Pál, C. Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. en. *Nat Microbiol* **3**, 718–731 (June 2018).
45. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (Feb. 2004).
46. Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñoz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I. & Karp, P. D. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. en. *Nucleic Acids Res.* **45**, D543–D550 (Jan. 2017).
47. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. en. *Nucleic Acids Res.* **44**, D286–93 (Jan. 2016).

48. The Gene Ontology Consortium & The Gene Ontology Consortium. *The Gene Ontology Resource: 20 years and still GOing strong* 2019.
49. Santos-Zavaleta, A., Sánchez-Pérez, M., Salgado, H., Velázquez-Ramírez, D. A., Gama-Castro, S., Tierrafria, V. H., Busby, S. J. W., Aquino, P., Fang, X., Pálsson, B. O., Galagan, J. E. & Collado-Vides, J. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. en. *BMC Biol.* **16**, 91 (Aug. 2018).
50. Turnbough Jr, C. L. & Switzer, R. L. Regulation of pyrimidine biosynthetic gene expression in bacteria: repression without repressors. en. *Microbiol. Mol. Biol. Rev.* **72**, 266–300, table of contents (June 2008).
51. Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. en. *Mol. Cell. Proteomics* **5**, 144–156 (Jan. 2006).
52. Ahrné, E., Molzahn, L., Glatzer, T. & Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **13**, 2567–2578 (2013).
53. Gansner, E. R. & North, S. C. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exp.* **30**, 1203–1233 (2000).
54. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Chapter 4

Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility

Responses of bacteria to antibiotic treatments depend on their environment. Differences between *in vitro* testing conditions and the physiological environments inside patients have resulted in poor antibiotic susceptibility predictions, contributing to treatment failures in the clinic. Here, we investigate how media composition affects antibiotic susceptibility, and contextualize these changes through machine learning of transcriptomics data. We show that complex transcriptional changes induced by different media or antibiotic treatment can be traced back to a few key regulators. Integration of results from machine learning with biochemical knowledge reveals fundamental shifts in respiration and iron availability that may explain media-dependent differential susceptibility to antibiotics. The data generation and analytical workflow used here

can interrogate the regulatory state of a pathogen under any condition, and can be extended to additional strains and organisms for which data is available.

4.1 Background

Antibiotic activities are highly dependent on environmental conditions. Factors such as metabolites, buffering systems, immune components, and the presence of other antibiotics can drastically alter their activities against bacteria, as measured by changes in their minimum inhibitory concentrations (MIC) [1–5]. Despite these dependencies, standardized antibiotic susceptibility testing (AST) is performed in bacteriological media such as Cation-Adjusted Mueller Hinton Broth (CA-MHB) which poorly recapitulates the *in vivo* conditions of the host [5]. AST in tissue culture media (e.g. RPMI, DMEM) more closely mimics physiological conditions and may be a better predictor of *in vivo* efficacy [5, 6]. Using transcriptomic data, we sought to understand the drivers of differential antibiotic susceptibility in these media conditions.

Prior studies have demonstrated the utility of systems analysis in global metabolic responses to antibiotic treatment using various omics measurements [7, 8]. Previously, we published a high-quality transcriptomic compendium for *Escherichia coli*, named PRECISE, containing 278 expression profiles [9]. Using independent component analysis (ICA), a machine learning method developed to separate mixed signals, we decomposed PRECISE into 92 independently modulated groups of genes, or i-modulons. Since most i-modulons exhibited high overlap with previously reported regulons, we named the i-modulons after their common regulator(s). ICA simultaneously computes activities for each i-modulon across every condition in the expression compendium, which predictably reflect the activity state of the enriched regulators.

I-modulon activities have been used to infer the cellular state beyond the transcriptome.

In one case, adaptive laboratory evolution of *E. coli* in excess iron conditions resulted in multiple mutations in the oxidative stress response regulator OxyR [10]. In each of these strains the SOS-response regulator LexA was no longer activated by hydrogen peroxide treatment, indicating that the mutations reduced DNA damage from reactive oxygen species through constitutive activation of OxyR. I-modulon activities related to the ferric uptake regulator Fur have also provided information on the intracellular iron concentration [11, 12].

In the present study, we demonstrate the utility of an ICA-based workflow in decomposing the transcriptomic responses to antibiotics in the *E. coli* K-12 laboratory strain MG1655. First, we show that the MIC of *E. coli* differs across three media compositions: (1) glucose M9 minimal media, a standard media for scientific discovery; (2) cation adjusted mueller hinton broth, the standard rich media for antibiotic susceptibility testing, and (3) RPMI 1640 with 10% LB, a rich media that mimics physiological conditions. We then measured the expression profiles of *E. coli* on the three media and identified key transcriptomic differences through ICA. Finally, we examined the expression profiles of *E. coli* after treatment with subinhibitory concentrations of antibiotics and used ICA to infer a deceleration of respiration in rich media. Overall, this work provides a promising avenue for interrogating the effects of antibiotic treatment in different environmental conditions.

4.2 Results

4.2.1 Antibiotic efficacy varies across media compositions

Antibiotic activities vary across media compositions We examined the MICs of nine antibiotics, ranging in drug class and mechanism of action, on glucose M9 minimal media (M9),

RPMI 1640 + 10% LB (R10LB) and cation-adjusted Mueller Hinton Broth (CA-MHB) (Table 1). Based on their activity profiles, we selected four antibiotics for further analysis: the *beta*-lactams Ceftriaxone (CTR) and Meropenem (MEM), the fluoroquinolone ciprofloxacin (CIP) , and the sulfonamide combination trimethoprim-sulfamethoxazole (T/S). Ceftriaxone and T/S had identical MICs in the rich media (CA-MHB and R10LB) and a lower MIC in M9. The other two antibiotics exhibited opposite susceptibility patterns across the three media types; as the lowest MICs for ciprofloxacin were in R10LB and in CA-MHB for meropenem.

Table 4.1: MIC of nine antibiotics on three media. Antibiotics in bold were selected for further analysis.

	Drug Information		MIC ($\mu\text{g/mL}$)		
	Drug Class	Target Process	M9	R10LB	CA-MHB
Ampicillin (AMP)	Penicillin	Cell Wall	>16	256	64
Ceftriaxone (CTR)	Cephalosporin	Cell Wall	0.016	0.25	0.25
Ciprofloxacin (CIP)	Quinolone	DNA Gyrase	0.008	0.004	0.016
Fosfomycin	Fosfomycin	Cell Wall	64	>64	16
Meropenem (MEM)	Carbapenem	Cell Wall	0.031	0.13	0.016
Nitrofurantoin	Furan	Multiple	128	>512	256
Plazomicin	Aminoglycoside	30S Inhibitor	0.13	0.13	0.25
Rifampicin	Rifampin	Transcription	8		16
Trimethoprim-sulfamethoxazole	Pyrimidine/Sulfa	Folate Synthesis	0.016	0.125	0.125

The MICs were measured in 96-well plates using an inoculum density of 10^5 CFU/mL, as recommended by the Clinical and Laboratory Standards Institute (CLSI). For RNA-sequencing, we required a significantly higher cell density ($\sim 5 * 10^7$ CFU/mL) to collect enough mRNA for sequencing. Since antibiotic activity can be diluted at higher cell densities, we determined the minimum bactericidal concentration (MBC) in each media type for the targeted antibiotics at the same culture density that was used for the mRNA isolations (Figure 4.1a). In most conditions, the MBC was higher than the selected MICs. We therefore measured the change in MIC at

RNA-seq densities, and found that the MICs were increased up to 8-fold at higher densities (Figure 4.1b).

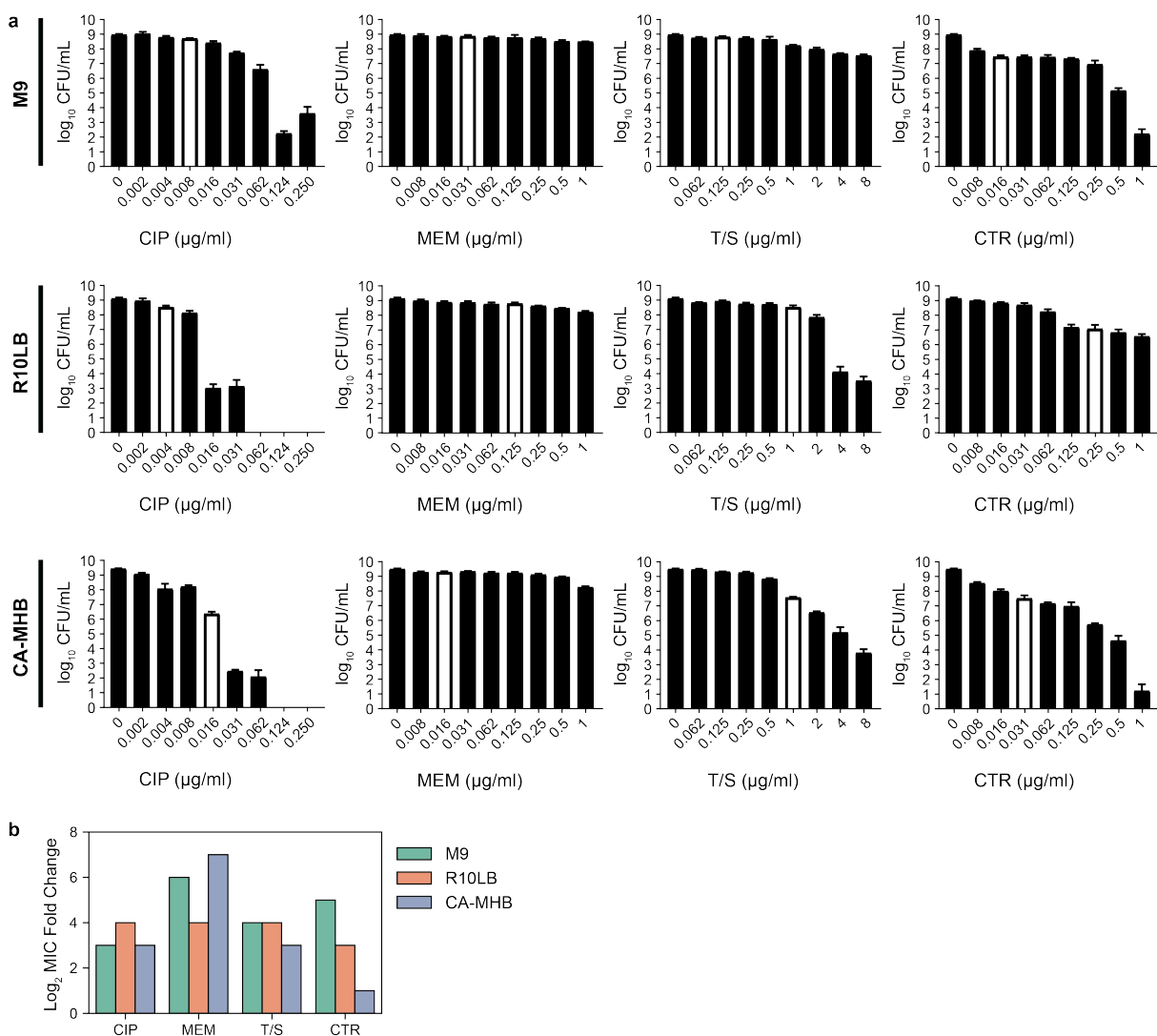


Figure 4.1: Inoculum cell density affects antibiotic efficacy. (a) MBC measurements of four antibiotics across three media compositions at gene expression conditions. White bars designate the concentration of antibiotics used in RNA-seq cultures (i.e. MIC in relevant media) (b) Changes in MIC between low ($\sim 10^5$ CFU/mL) and high ($\sim 10^7$ CFU/mL) cell densities.

4.2.2 A few mechanisms can explain large transcriptomic perturbations

We first performed RNA-seq on the untreated strains grown in M9, R10LB, and CA-MHB to understand how *E. coli* responds to different media without antibiotic treatment. Pairwise differential expression analysis of the three growth conditions identified up to 1145 differentially expressed genes (DEGs) between growth conditions, comprising one quarter of all genes in *E. coli* (Figure 1a). The DEGs spanned every Cluster of Orthologous Gene (COG) category, with no known function for 21% of all DEGs. Rich media (R10LB and CA-MHB) induced similar transcriptional responses, as 713 genes were commonly differentially expressed in either rich media when compared to M9 minimal media.

The acquired RNA-seq data was compared to a previously published high-quality compendium containing 278 *E. coli* expression profiles, named PRECISE, to contextualize the transcriptional response to the three media types [9]. Principal component analysis (PCA) of the combined dataset showed that CA-MHB and R10LB produced responses that were significantly different than the previously probed conditions in PRECISE, including growth on LB media without RPMI (Figure 4.2b). On the other hand, the expression levels on M9 minimal media were highly similar to expression levels on an identical condition produced five years ago ($R^2 = 0.85$) [13], highlighting the internal consistency of this dataset. The older expression profile grown on glucose minimal media serves as the reference condition, to which all gene expression levels are compared.

We previously showed that independent component analysis (ICA) can be applied to large RNA-seq datasets to extract independently modulated sets of genes, termed i-modulons, and their condition-specific activities [9]. I-modulons exhibit high overlap with classically-defined regulons, which allows us to directly associate i-modulons to one or more known transcriptional

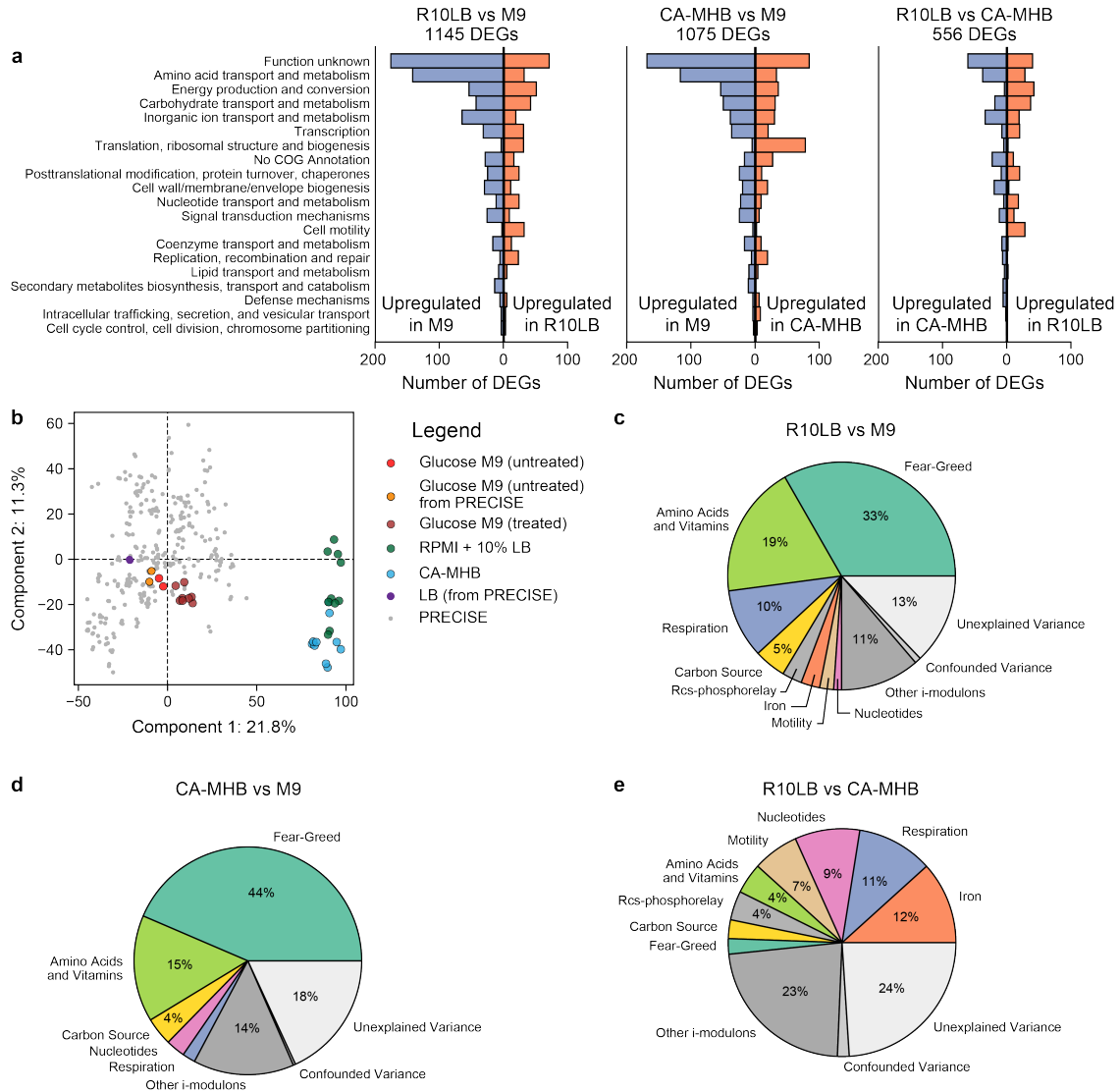


Figure 4.2: Global comparative analysis of the *E. coli* transcriptome across three media compositions. (a) Differentially expressed genes across COG categories between three media compositions. (b) Top two principal component loadings for the PRECISE dataset, comprised of 279 RNA-seq profiles [9], combined with the data generated in this study. (c-f) Proportion of expression deviation explained by 8 groups of i-modulons. I-modulons in each group are shown in Figure 3 and Figure S1. Percentages are shown when above 3%.

regulators. Unlike many module detection methods [14], ICA also computes activity levels for each i-modulon across all expression profiles, indicating the activity level of the linked regulator under each specific condition.

We applied ICA to PRECISE combined with 30 new RNA-seq datasets to form a compendium of 308 profiles and identified 98 i-modulons, of which 88 were nearly identical to i-modulons extracted from the original PRECISE dataset of 278 profiles [9]. On average, i-modulons could explain 81% of the difference in expression between each pair of media conditions (Figure 4.2c-e). Borrowing from multiple linear regression, we also calculated how much of the expression deviation could be directly explained by each i-modulon (See Methods).

4.2.3 Transcriptional markers of ‘greedy’ growth in rich media

The major source of expression variation between either rich media and minimal media was explained by the ‘fear-greed’ tradeoff [9] (Figure 4.3a). This trade-off is between fast growth (i.e. ‘greed’) and high stress response readiness (i.e. ‘fear’). In rich media, cells are significantly less stressed by their environment, leading to a strong decrease in the activity level of stress-related i-modulons, such as RpoS and the GadEWX acid-response system [15]. In turn, this frees cellular resources that can be allocated to promote growth, as indicated by the increases in activity levels of i-modulons encoding genes related to translation and the central dogma of molecular biology. These two processes are molecularly linked through ppGpp, the stress alarmone, which binds to RNA polymerase to wobble the transcriptome [16]. The i-modulon activities under rich media are similar to the i-modulon activities of an RpoB mutant acquired during adaptive laboratory evolution on glucose minimal media [17], which is hypothesized to affect the ppGpp binding site on RNA polymerase.

R10LB contains glucose in its formulation, but surprisingly shows high transcriptomic similarities to glucose-starved media (Figure 4.3b). The cAMP receptor protein, CRP, is activated when glucose is absent from the media, recruiting RNA polymerase to genes encoding catabolic

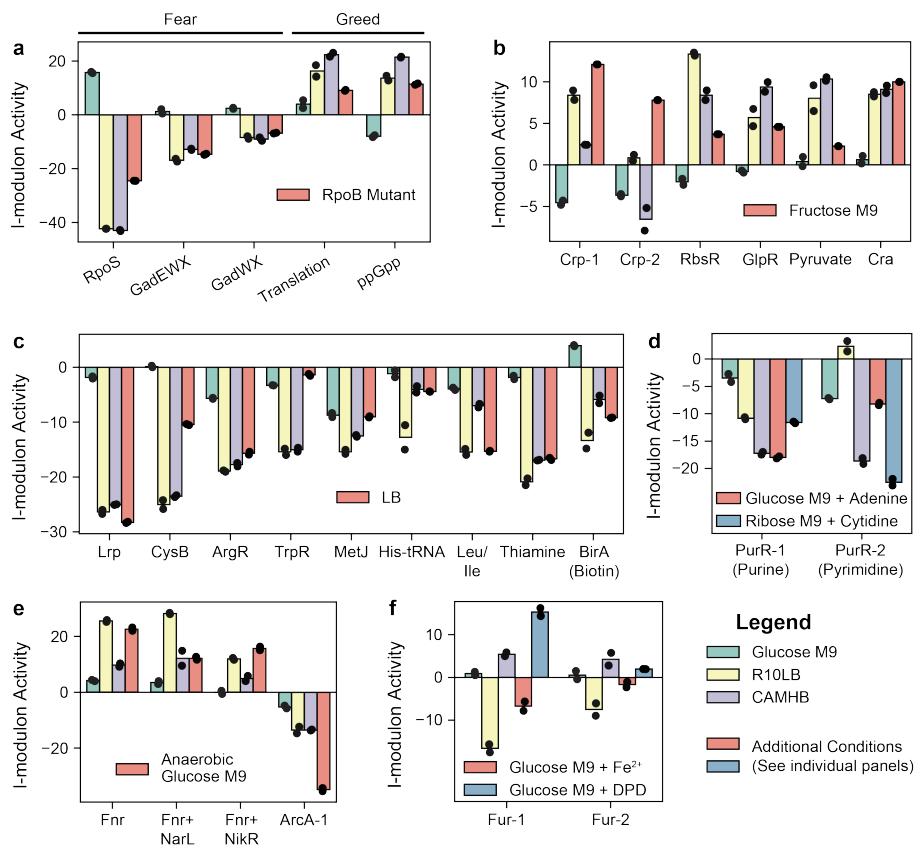


Figure 4.3: Mechanisms underlying the complex transcriptional response to different media. Bar charts show the i-modulon activities for (a) fear-greed i-modulons; (b) carbon source catabolism i-modulons; (c) amino acid and vitamin B i-modulons; (c) nucleotide biosynthesis i-modulons; (e) respiration i-modulons; (f) iron-related i-modulons. Individual measurements for biological replicates are plotted on top of bars.

pathways for alternative carbon sources. Although it is known that CA-MHB does not contain glucose, some of its CRP-related i-modulon activities are closer to M9 media than R10LB.

Biosynthetic pathways for vitamins and amino acids contributed 19% and 15% of the expression deviation from M9 for R10LB and CA-MHB, respectively (Figure 4.3c). These responses are consistent with the defined formulation of RPMI, which contains all amino acids and multiple B-vitamins, and shows that the undefined media CA-MHB and LB have a similar nutritional content to R10LB. The main deviations appear to be in histidine and branched-chain amino acid biosynthesis, which are all regulated by tRNA-mediated transcriptional attenuation. The

two media also deviate with respect to nucleotides as pyrimidine biosynthesis is upregulated in R10LB, indicating that cells are starved for pyrimidines (Figure 4.3d). Although nucleotides are not in the RPMI formulation, the cells downregulated the de novo purine biosynthesis pathway.

The next largest contributor to expression deviation in R10LB included four i-modulons related to anaerobic respiration (Figure 4.3e). Three i-modulons were enriched with the transcription factor Fnr; one i-modulon was uniquely enriched with the Fnr regulon, one i-modulon was enriched with genes co-regulated by Fnr and the nitrate-responsive regulator NarL and the final i-modulon contained genes co-regulated by Fnr, NarL, and the nickel regulator NikR. Fnr contains an iron-sulfur cluster that is destabilized in the presence of molecular oxygen [18], rendering Fnr inactive. Therefore, the high expression of these genes in R10LB indicates a high respiration rate that leaves little intracellular molecular oxygen available to destabilize Fnr's iron-sulfur cluster. Herein lies one of the major differences between the two media; R10LB promotes a higher respiration rate than CA-MHB. The fourth i-modulon in this group was related to ArcAB, a two-component system that senses the redox state of quinone electron carriers to regulate the TCA cycle, among other genes [19, 20]. The ArcA i-modulon activity was partially reduced in both R10LB and CA-MHB.

Another major difference between the two media is the presence of free iron, as indicated by the differential activation of two Fur-enriched i-modulons (Figure 4.3f). Fur binds to free iron (II) and represses genes related to iron siderophore synthesis and transport [21]. From the i-modulon activities, cells growing in glucose M9 media contain more free iron than CA-MHB, but significantly less than R10LB.

A few additional i-modulons were differentially activated across the multiple media conditions, and coincidentally discriminate the biological processes that are perturbed between the old

and new glucose M9 expression profiles. The new glucose M9 and CA-MHB expression profiles displayed an activation of three i-modulons related to the Rcs-phosphorelay and membrane stress (Figure D.1a), whereas the R10LB i-modulon activities were near the reference condition. The FliA i-modulon, which controls chemotaxis, was activated in all three media, whereas the FlhDC i-modulon, which controls biosynthesis and assembly of flagella, was strongly activated in RPMI (Figure D.1b).

4.2.4 Subinhibitory antibiotic treatment decelerates respiration

Since the ICA-based decomposition of transcriptomes provided a clear explanation of the complex responses to different media compositions, we used a similar approach to investigate the effects of antibiotic treatment at the subinhibitory concentrations presented in Figure 4.1. The i-modulon activities for the antibiotic-treated conditions were primarily clustered by media (Figure D.2), even when controlling for the media-specific responses (Figure 4.4a). Meropenem treatment on R10LB had minimal effects on the transcriptome and was removed from subsequent analyses.

We first investigated common trends in i-modulon activities upon treatment by any antibiotic. Antibiotic treatment increased activity of the translation i-modulon in all media compositions (Figure 4.4b), but left the other Fear-Greed trade-off i-modulons largely unchanged (Figure D.3). Although none of the selected antibiotics directly interact with the ribosome, this indicated that antibiotic-mediated stress could indirectly increase ribosome levels in the cell.

Across PRECISE, we observed a strong significant negative correlation between Fnr and ArcA i-modulon activities (Pearson $R = -0.74$, $p\text{-value} < 10^{-10}$), which represent two major respiration-related transcription factors (Figure 4.4c). In both rich media, there were significant

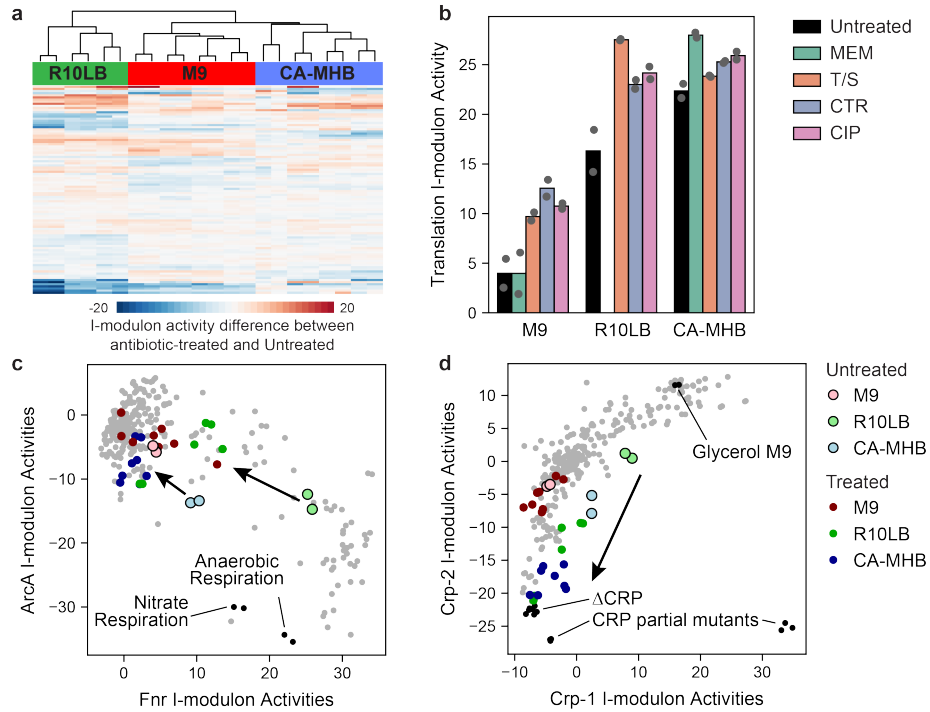


Figure 4.4: Consistent i-modulon responses to antibiotic treatment. (a) Clustered heatmap of expression differences in antibiotic-treated samples. Expression differences shown are between the antibiotic treated sample and the untreated sample in the same media condition. (b) Bar chart of translation i-modulon activities with and without antibiotic treatment. (c) Scatter plot of Fnr and ArcA i-modulon activities across PRECISE. Arrows indicate the effect of antibiotic treatment in rich media. Anaerobic respiration conditions from PRECISE are also shown. (d) Scatter plot of the Crp-1 and Crp-2 i-modulon activities across PRECISE. Arrow indicates the effect of antibiotic treatment in rich media. I-modulon activities from CRP knock-out and partial knock-out strains are also shown.

drops in Fnr i-modulon activities, indicating higher intracellular oxygen levels. This drop was coupled by a simultaneous increase in ArcA-regulated genes in the TCA cycle. Together, these two i-modulons point towards deceleration of the high respiratory rate facilitated by the rich media and a reduction in the redox state, as sensed by the ArcAB response system [19]. A prior study used biochemical assays to investigate the effects of bacteriostatic antibiotics on *E. coli* [22], and observed that antibiotic treatment drastically reduced the respiration rate and moderately reduced the redox state of the cell. However, this effect seemed to be mitigated in minimal media, where cellular respiration was initially slower.

We also observed significant reduction in both Crp-related i-modulon activities upon antibiotic treatment in rich media (Figure 4.4d). The Crp i-modulons encode genes in secondary carbon source catabolism; the Crp-1 i-modulon contains scavenging enzymes such as tryptophanase and acetyl-coA synthetase, whereas the Crp-2 i-modulon represents expression of phosphotransferase systems for alternative carbon sources.

Nutrient usage in the two rich media was nearly identical in the untreated samples, except for i-modulons encoding histidine biosynthesis (His-tRNA), branched-chain amino acid biosynthesis (Leu/Ile), and nucleotide biosynthesis (PurR-1 and PurR-2). However, upon treatment by any antibiotic, the activities of these i-modulons converged to similar levels in rich media (Figure D.4).

4.2.5 Ciprofloxacin has diverging effects on Fur i-modulons

For most i-modulons, treatment by any antibiotic in subinhibitory concentrations resulted in a convergence of i-modulon activity between R10LB and CA-MHB. However, these mirrored responses do not explain why drugs like ciprofloxacin or meropenem have diverging MICs on the two media. Here, we will focus on ciprofloxacin and suggest a hypothesis towards its differential activity on the two media.

Ciprofloxacin blocks DNA gyrase, creating double-stranded DNA breaks [23]. Subsequently, we observed a strong increase in LexA activity in all three media (Figure 4.5a). In addition, ciprofloxacin has diverging effects on the two Fur-related i-modulons. The two Fur i-modulons contain considerable overlap (Figure S5), and exhibit a general trend in their activities across the entire PRECISE database (Figure 4.5b). The Fur-1 i-modulon responds to iron starvation, de-repressing iron (II) and (III) siderophore synthesis and transport systems,

ribonucleotide reductases, superoxide dismutase, and iron-sulfur cluster assembly. The Fur-2 i-modulon responds to excess iron, further repressing siderophore transport and repressing the energy-transducing Ton system.

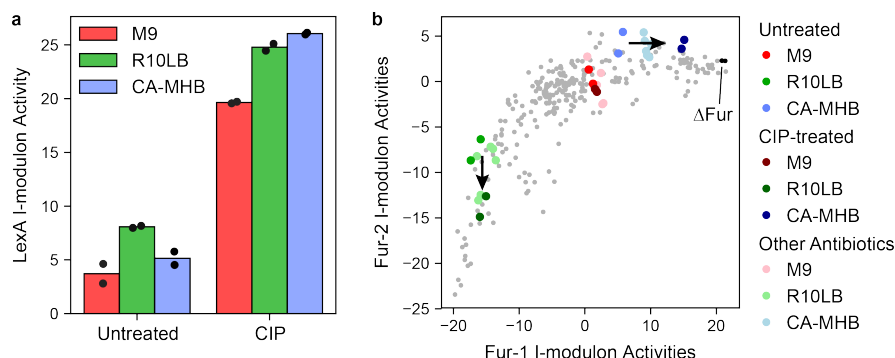


Figure 4.5: I-modulon responses to ciprofloxacin treatment. (a) The DNA-damage response regulator LexA is activated by ciprofloxacin treatment. (b) Scatter plot of the Fur-1 and Fur-2 i-modulon activities. Expression profiles from PRECISE are shown in light gray, and other colors are described in the legend. Arrows indicate i-modulon shifts in R10LB and CA-MHB resulting from ciprofloxacin treatment.

The Fur i-modulon activities diverge further upon ciprofloxacin treatment (Figure 5b), indicating an increase in free iron in R10LB and a stronger iron starvation response in CA-MHB. [what could cause this]. Between the increased intracellular oxygen levels observed in the previous section, and the increase in free iron specific to R10LB, the increased susceptibility of *E. coli* to ciprofloxacin in R10LB may be explained by Fenton chemistry, in which unincorporated iron and hydrogen peroxide generate hydroxyl radicals to further damage DNA [24]. Prior studies have observed that reactive oxidative species, such as those created by Fenton chemistry, may contribute to the antibacterial action of ciprofloxacin [25–27]. However, we did not observe a change in activity for any oxidative response regulons (OxyR, SoxR, and RpoS), potentially due to cell density effects.

4.3 Discussion

Bacterial responses to antibiotics are complex, and depend on a variety of environmental and genetic factors. In this study, we explored the effects of different media compositions and antibiotic treatments on the transcriptomic state of the model organism *E. coli*. First, we showed that i-modulons could simplify the interpretation of large changes in expression due to varying environmental conditions. We then leveraged i-modulon activities with biochemical knowledge to infer the respiratory rate of *E. coli*, and subsequently elucidate a drop in respiration caused by antibiotic treatment. We also noted a significant difference in Fur activity between the two rich media, providing a potential hypothesis behind the differential susceptibility to ciprofloxacin in the two media.

I-modulons played a central role in contextualizing the gene expression changes. The use of ICA enabled a rapid, low-dimensional characterization of one thousand differentially expressed genes, many of which were uncharacterized. In addition, biochemical knowledge of i-modulons provided connections between the transcriptome and intracellular states, such as respiration rates and biosynthetic pathways. However, it is not clear if iron levels directly affect ciprofloxacin, especially in light of the controversy regarding the role of oxidative stress in antibiotic lethality [28–30].

We have presented an analytical pipeline for probing and contextualizing the transcriptomic effects of antibiotic treatment using a model organism. Although our analysis was focused on subinhibitory effects of antibiotics, this could be extended to investigate inhibitory or lethal doses of antibiotics. As publicly available transcriptomic datasets grow in size and diversity [31], this pipeline can be applied to pathogenic strains and organisms grown in physiologically relevant media conditions. Future studies could explore the metabolic and transcriptional shifts

associated with antibiotic treatment, leading to a “white-box” approach that connects antibiotic-treated transcriptomes to biochemical knowledge [8].

4.4 Methods

4.4.1 Bacterial strains and growth conditions

Escherichia coli strain MG1655 was grown in three media: (1) M9 minimal medium (47.8 mM Na₂HPO₄, 22 mM KH₂PO₄, 8.6 mM NaCl, 18.7 mM NH₄Cl, 2 mM MgSO₄ and 0.1 mM CaCl₂) supplemented with 0.2% (w/v) glucose (M9); (2) Roswell Park Memorial Institute 1640 (RPMI) (Thermo Fisher Scientific) supplemented with 10% LB (R10LB); (3) Mueller Hinton Broth (Sigma-Aldrich) supplemented with 25 mg/L Ca²⁺ and 12.5 mg/L Mg²⁺ (CA-MHB). Glycerol stocks of *E. coli* were inoculated into the respective media and cultured overnight at 37C with vigorous agitation.

4.4.2 Antibiotic susceptibility testing

Antibiotic susceptibility was determined as described previously [32]. Briefly, *E. coli* were cultured in the same media throughout (M9, R10LB, or CA-MHB) prior to the addition of antibiotics. Mid-logarithmic phase cultures were diluted to approximately $\sim 2.5 \times 10^7$ CFU/mL (OD₆₀₀ = 0.08) and diluted 1:100 in 200 μ L media prepared with serial dilution. Plates were incubated with shaking at 200 rpm at 37°C overnight. Bacterial growth (OD₆₀₀) was determined approximately 20 h later utilizing a Enspire Alpha multimode plate reader (PerkinElmer). To calculate the MIC₉₀, defined as the amount of drug required to inhibit $\leq 90\%$ of the growth of the untreated controls, the density of each drug-treated well was compared to untreated control. MIC for high-density cultures were performed as described above, except with a starting OD₆₀₀

~ 0.05 after dilution.

4.4.3 Bactericidal Assays

Mid-log *E. coli* MG1655 was used to inoculate 15 ml of either glucose M9, R10LB, or CA-MHB with approximately 2.5×10^7 CFU/mL (OD600 ~ 0.05). Each experimental well of the 96-well flat bottom plate received 180 μ l of bacterial culture and 20 μ l of the desired 10x drug stock. Plates were incubated shaking at 100 rpm at 37C overnight. After 20 h, plates were removed from the incubator and serial 10-fold dilutions of each well performed in their respective media. Twenty microliters of each serial dilution was spot plated onto LA and incubated at 37C overnight to enumerate the CFU.

4.4.4 RNA-seq expression profiling and processing

Total RNA was sampled from duplicate cultures. Strains were cultured overnight in the respective media as described above. Mid-log phase cultures were diluted (starting OD600 ~ 0.05) into 30mL media for untreated conditions, or 30 mL media containing the MIC of the respective antibiotic for the respective media (Table 1). Flasks were incubated for 30 min at 37C with shaking. Cell broth was centrifuged and supernatant was removed. RNA extraction and library preparation were performed as described in [9]. Raw sequencing reads were performed as described in [9]. Differential expression was performed using DESeq2 [33], with a log2 fold change cutoff of 1.5 and q-value cutoff of 0.05.

4.4.5 Independent component analysis

Log-transformed transcripts per million (log-TPM) expression levels were concatenated to the PRECISE dataset. ICA and i-modulon extraction were performed as described in [9].

Acknowledgements

The authors would like to thank Dr. George Sakoulas, Hannah Tsunemoto, and Yara Seif for informative discussions. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was funded by the Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (grant number NNF10CC1016517) and the NIH NIAID grant (1-U01-AI124316)

Author Contributions: A.V.S., N.D., A.M.F., and B.O.P. designed the study. Y.H., S.X., and R.S. performed MIC testing and RNA-seq preparation. N.D. performed killing assays. A.V.S. and S.P. performed computations and interpretation of results. A.V.S., S.P. and B.O.P. wrote and edited the manuscript.

Chapter 4, in part, is currently being prepared for submission for publication. **AV Sasttry***, N Dillon*, S Poudel, Y Hefner, S Xu, R Szubin, AM Feist, BO Palsson “Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility”. The dissertation author is one of the two primary authors.

4.5 References

1. Meylan, S., Andrews, I. W. & Collins, J. J. Targeting Antibiotic Tolerance, Pathogen by Pathogen. en. *Cell* **172**, 1228–1238 (Mar. 2018).
2. Farha, M. A., French, S., Stokes, J. M. & Brown, E. D. Bicarbonate Alters Bacterial Susceptibility to Antibiotics by Targeting the Proton Motive Force. *ACS Infect. Dis.* **4**, 382–390 (Mar. 2018).
3. Dorschner, R. A., Lopez-Garcia, B., Peschel, A., Kraus, D., Morikawa, K., Nizet, V. & Gallo, R. L. The mammalian ionic environment dictates microbial susceptibility to antimicrobial defense peptides. *The FASEB Journal* **20**, 35–42 (Jan. 2006).

4. Kudrin, P., Varik, V., Oliveira, S. R. A., Beljantseva, J., Del Peso Santos, T., Dzhygyr, I., Rejman, D., Cava, F., Tenson, T. & Hauryliuk, V. Subinhibitory Concentrations of Bacteriostatic Antibiotics Induce *relA*-Dependent and *relA*-Independent Tolerance to β -Lactams. en. *Antimicrob. Agents Chemother.* **61** (Apr. 2017).
5. Ersoy, S. C., Heithoff, D. M., Barnes, L., Tripp, G. K., House, J. K., Marth, J. D., Smith, J. W. & Mahan, M. J. Correcting a Fundamental Flaw in the Paradigm for Antimicrobial Susceptibility Testing. *EBioMedicine* **20**, 173–181 (June 2017).
6. Lin, L., Nonejuie, P., Munguia, J., Hollands, A., Olson, J., Dam, Q., Kumaraswamy, M., Rivera, H., Corriden, R., Rohde, M., Hensler, M. E., Burkart, M. D., Pogliano, J., Sakoulas, G. & Nizet, V. *Azithromycin Synergizes with Cationic Antimicrobial Peptides to Exert Bactericidal and Therapeutic Activity Against Highly Multidrug-Resistant Gram-Negative Bacterial Pathogens* 2015.
7. Zampieri, M., Zimmermann, M., Claassen, M. & Sauer, U. Nontargeted Metabolomics Reveals the Multilevel Response to Antibiotic Perturbations. en. *Cell Rep.* **19**, 1214–1228 (May 2017).
8. Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C. & Collins, J. J. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. en. *Cell* **177**, 1649–1661.e9 (May 2019).
9. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. en. *Nat. Commun.* **10**, 5536 (Dec. 2019).
10. Anand, A., Chen, K., Catoiu, E., Sastry, A. V., Olson, C. A., Sandberg, T. E., Seif, Y., Xu, S., Szubin, R., Yang, L., Feist, A. M. & Palsson, B. O. OxyR is a convergent target for mutations acquired during adaptation to oxidative stress-prone metabolic states. en. *Mol. Biol. Evol.* (Oct. 2019).
11. Anand, A., Olson, C. A., Yang, L., Sastry, A. V., Catoiu, E., Choudhary, K. S., Phaneuf, P. V., Sandberg, T. E., Xu, S., Hefner, Y., Szubin, R., Feist, A. M. & Palsson, B. O. Pseudogene repair driven by selection pressure applied in experimental evolution. en. *Nat Microbiol* **4**, 386–389 (Mar. 2019).
12. Anand, A., Chen, K., Yang, L., Sastry, A. V., Olson, C. A., Poudel, S., Seif, Y., Hefner, Y., Phaneuf, P. V., Xu, S., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. en. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 25287–25292 (Dec. 2019).
13. Kim, D., Seo, S. W., Gao, Y., Nam, H., Guzman, G. I., Cho, B.-K. & Palsson, B. O. Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. en. *Nucleic Acids Res.* **46**, 2901–2917 (Apr. 2018).
14. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. en. *Nat. Commun.* **9**, 1090 (Mar. 2018).

15. Seo, S. W., Kim, D., O'Brien, E. J., Szubin, R. & Palsson, B. O. Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in *Escherichia coli*. en. *Nat. Commun.* **6**, 7970 (Aug. 2015).
16. Sanchez-Vazquez, P., Dewey, C. N., Kitten, N., Ross, W. & Gourse, R. L. Genome-wide effects on *Escherichia coli* transcription from ppGpp binding to its two sites on RNA polymerase. en. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8310–8319 (Apr. 2019).
17. Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. en. *Cell Syst* **2**, 260–271 (Apr. 2016).
18. Kiley, P. J. & Beinert, H. Oxygen sensing by the global regulator, FNR: the role of the iron-sulfur cluster. en. *FEMS Microbiol. Rev.* **22**, 341–352 (Dec. 1998).
19. Georgellis, D., Kwon, O. & Lin, E. C. Quinones as the redox signal for the arc two-component system of bacteria. en. *Science* **292**, 2314–2316 (June 2001).
20. Federowicz, S., Kim, D., Ebrahim, A., Lerman, J., Nagarajan, H., Cho, B.-K., Zengler, K. & Palsson, B. Determining the control circuitry of redox metabolism at the genome-scale. en. *PLoS Genet.* **10**, e1004264 (Apr. 2014).
21. Seo, S. W., Kim, D., Latif, H., O'Brien, E. J., Szubin, R. & Palsson, B. O. Deciphering Fur transcriptional regulatory network highlights its complex role beyond iron metabolism in *Escherichia coli*. en. *Nat. Commun.* **5**, 4910 (Sept. 2014).
22. Lobritz, M. A., Belenky, P., Porter, C. B. M., Gutierrez, A., Yang, J. H., Schwarz, E. G., Dwyer, D. J., Khalil, A. S. & Collins, J. J. Antibiotic efficacy is linked to bacterial cellular respiration. en. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8173–8180 (July 2015).
23. Chen, C. R., Malik, M., Snyder, M. & Drlica, K. DNA gyrase and topoisomerase IV on the bacterial chromosome: quinolone-induced DNA cleavage. en. *J. Mol. Biol.* **258**, 627–637 (May 1996).
24. Imlay, J. A. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. en. *Nat. Rev. Microbiol.* **11**, 443–454 (July 2013).
25. Goswami, M., Mangoli, S. H. & Jawali, N. Involvement of reactive oxygen species in the action of ciprofloxacin against *Escherichia coli*. en. *Antimicrob. Agents Chemother.* **50**, 949–954 (Mar. 2006).
26. Dwyer, D. J., Kohanski, M. A., Hayete, B. & Collins, J. J. Gyrase inhibitors induce an oxidative damage cellular death pathway in *Escherichia coli*. en. *Mol. Syst. Biol.* **3**, 91 (Mar. 2007).
27. Dwyer, D. J., Belenky, P. A., Yang, J. H., MacDonald, I. C., Martell, J. D., Takahashi, N., Chan, C. T. Y., Lobritz, M. A., Braff, D., Schwarz, E. G., Ye, J. D., Pati, M., Vercruysse, M., Ralifo, P. S., Allison, K. R., Khalil, A. S., Ting, A. Y., Walker, G. C. & Collins, J. J.

- Antibiotics induce redox-related physiological alterations as part of their lethality. en. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2100–9 (May 2014).
28. Keren, I., Wu, Y., Inocencio, J., Mulcahy, L. R. & Lewis, K. *Killing by Bactericidal Antibiotics Does Not Depend on Reactive Oxygen Species* 2013.
 29. Liu, Y. & Imlay, J. A. Cell death from antibiotics without the involvement of reactive oxygen species. en. *Science* **339**, 1210–1213 (Mar. 2013).
 30. Dwyer, D. J., Collins, J. J. & Walker, G. C. Unraveling the physiological complexities of antibiotic lethality. en. *Annu. Rev. Pharmacol. Toxicol.* **55**, 313–332 (2015).
 31. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (Jan. 2013).
 32. Dillon, N., Holland, M., Tsunemoto, H., Hancock, B., Cornax, I., Pogliano, J., Sakoulas, G. & Nizet, V. Surprising synergy of dual translation inhibition vs. *Acinetobacter baumannii* and other multidrug-resistant bacterial pathogens. en. *EBioMedicine* **46**, 193–201 (Aug. 2019).
 33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biol.* **15**, 550 (2014).

Chapter 5

Conclusions

Systems biology is now firmly in the realm of big data analytics. There is enough knowledge about specific biological parts that we can begin to understand biological systems as a whole. The inherent complexity of biological systems requires the production, integration, and processing of big omics datasets to accelerate our understanding of a living cell. In this thesis, we introduced independent component analysis as a promising tool to enable rapid knowledge generation from big transcriptomics data.

In the introduction, we illustrated the transcriptional regulatory network as a black box, where the inputs and outputs can be measured, but little is known about the internal mechanisms. However, ICA enables an interpretable white-box model for the TRN [1] that can explain the internal logic of the TRN (Figure 5.1).

We introduced the utility of ICA in Chapter 2 by applying it to PRECISE, which is currently the largest publicly-available RNA-seq compendium from a single research group for any bacteria. We have meticulously annotated each experimental condition, and believe that it will be a frequently used resource in the community. The 92 i-modulons extracted from

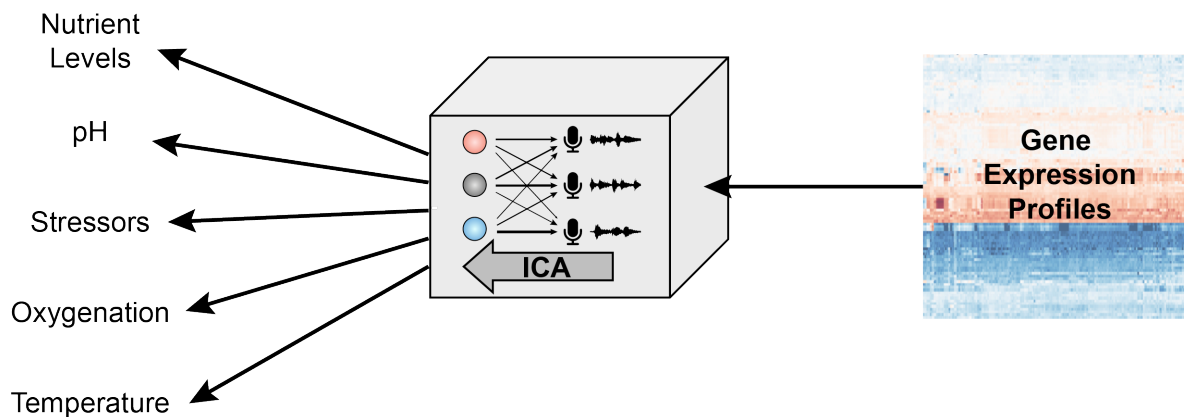


Figure 5.1: ICA presents a white-box model of the TRN

PRECISE revealed some of the deep knowledge accessed by ICA, including regulator discovery and genotype-phenotype relationships.

Chapter 3 expanded on the S matrix, which encodes the data-derived relationships between transcriptional regulators and genes. Through analysis of multiple independent transcriptomics and proteomics datasets, we demonstrated that the ICA-derived structure of the transcriptome is consistent across datasets and is robust to data integration. This chapter also highlighted that more work is required in identifying the optimal dimensionality of transcriptomics datasets, and introduced i-modulon coalescence as a property that can assist in this investigation.

Finally, we explored the implications of i-modulon activities in Chapter 4. I-modulons greatly simplified the analysis of complex transcriptional shifts, such as those induced by changes in media composition or treatment by antibiotics. Changes in i-modulon activity levels illustrated how ICA can extend the interpretation of transcriptomic datasets beyond the traditional limits of gene expression. Specifically, we showed that the Fnr i-modulon activities mimic the respiration rate of *E. coli*. This knowledge can be applied to future studies, highlighting the synergy unlocked by data integration.

Although the linear model of blind source separation does not accurately capture some intricate regulatory mechanics, such as regulator hierarchy or non-linear interactions, the investigations presented in this thesis demonstrate that ICA extracts accurate and interpretable information for bacterial transcriptomes. It is currently unclear if these assumptions are justifiable for higher organisms, but preliminary studies have shown promise [2–4]. However, ICA can be applied to other microbes, such as infectious diseases or members of the microbiome, to probe their responses to host environments or antibiotics.

Big data provides a comprehensive view of biological systems, but does not inherently provide any knowledge. Through this work, we have shown that ICA enables interpretation of big expression datasets to expose the hidden signals sensed by the cell.

Appendix A

The *E. coli* transcriptome mostly
consists of independently modulated
groups of genes - Supplementary
Information

A.1 Supplementary Figures

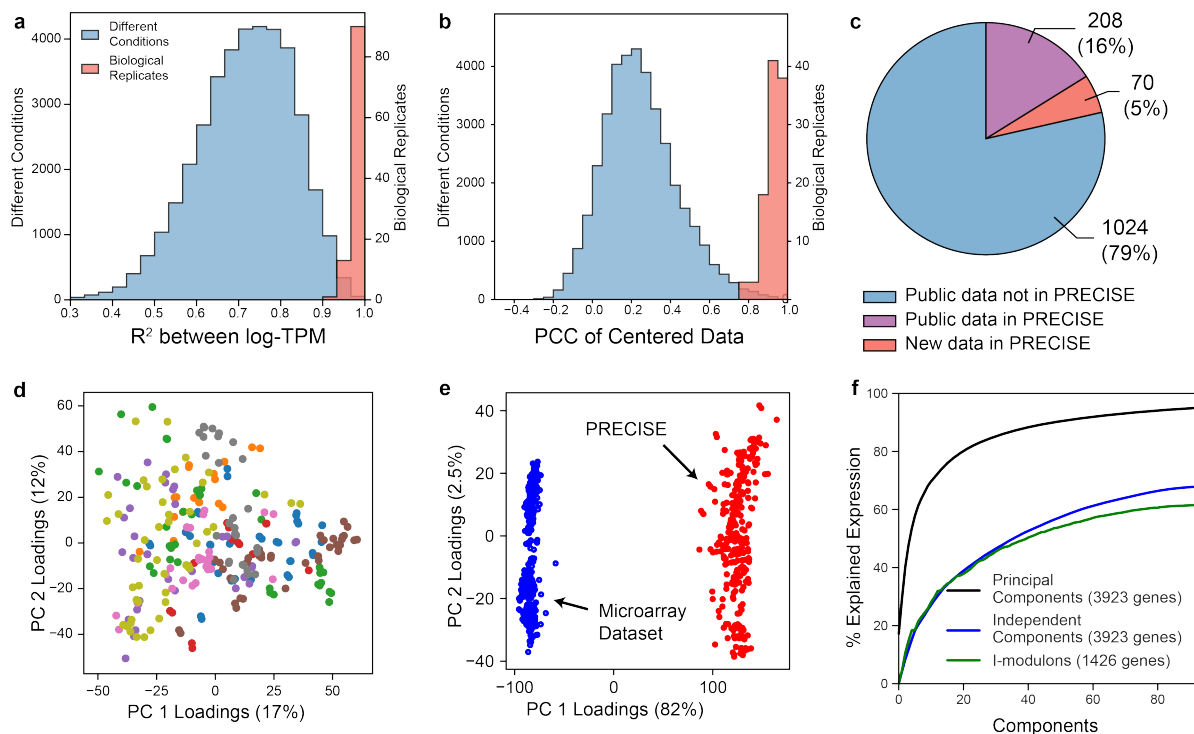


Figure A.1: Overview of the PRECISE Dataset. (a) Histogram of data quality, as measured by the coefficient of determination (R^2) between log-TPM. Comparison between biological replicates are shown in red, whereas comparison between all pairwise non-replicates are shown in blue. (b) Histogram of the Pearson correlation coefficient (PCC) between log-TPM, after subtraction of a reference condition (See Methods). (c) Pie chart of all published RNA-seq data for *E. coli* MG1655 or BW25113 on NCBI GEO as of January 2019. (d) Loadings of the first two principal components (PCs) of PRECISE, colored by the researcher who created the dataset. The high overlap of various colors demonstrates the self-consistency of the compendium. (e) Loadings of the first two principal component of PRECISE merged with a microarray dataset[1], illustrating the clear batch effects introduced by integrating data from different sources. (f) Cumulative explained variance for the first 92 components calculated by principal component analysis (black) and ICA (blue). Cumulative explained variance using only significant genes in each i-modulon is shown in green.

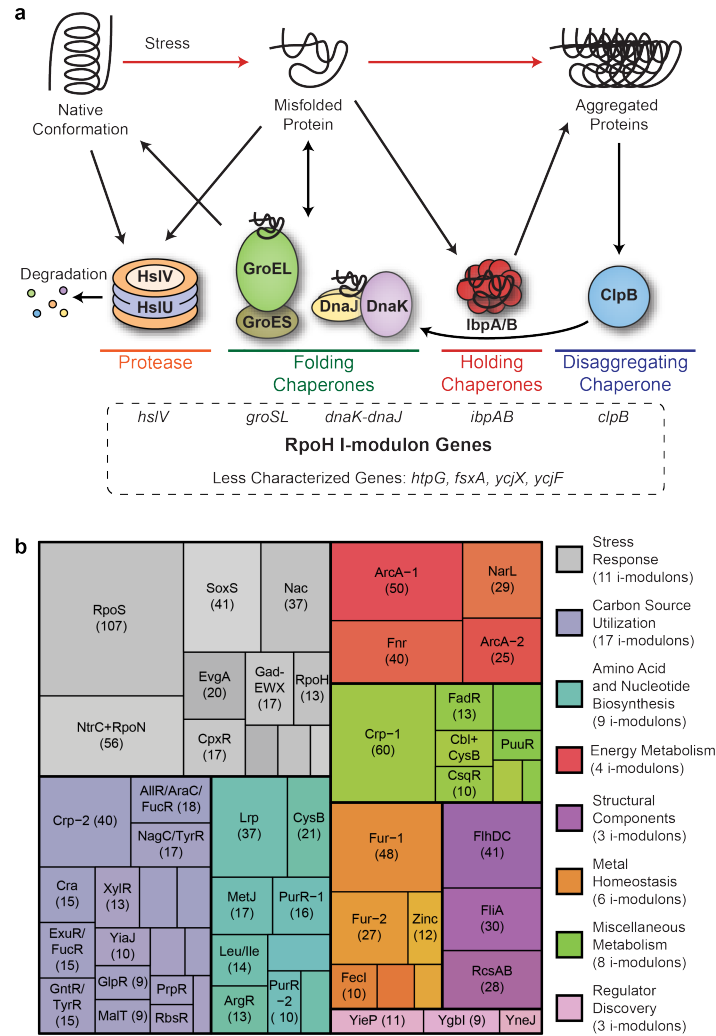


Figure A.2: Overview of i-modulons. (a) Simplified schematic illustration of the protein misfolding response in *E. coli* emphasizing genes in the RpoH i-modulon. Adapted from Baneyx and Mujacic, 2004 [2]. (b) Illustration of the 61 regulatory i-modulons, colored by functional category. Number of genes in each i-modulon are indicated by the size of the box.

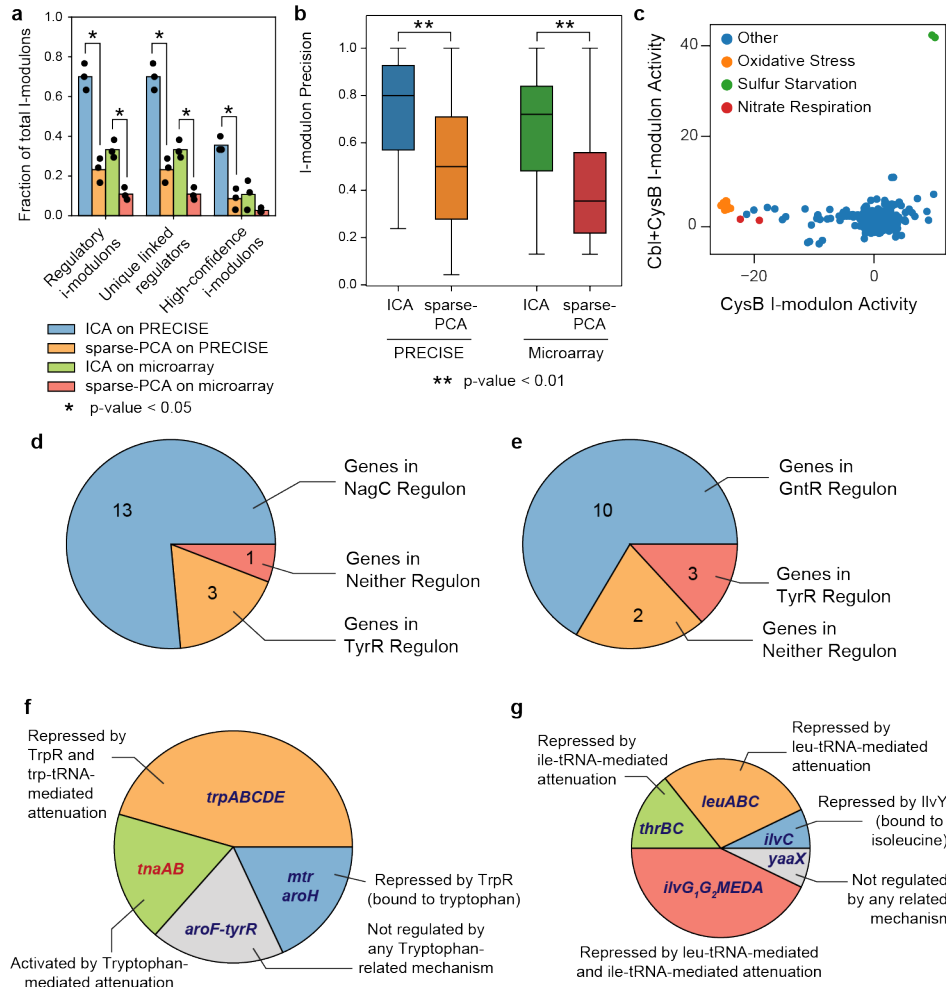


Figure A.3: Validation of i-modulon-regulator relationships. (a) Evaluation of ICA and sparse-PCA on PRECISE and a microarray dataset (NCBI GEO GSE6836)[1]. ICA consistently outperforms sparse-PCA (p -value < .05) in number of regulatory i-modulons, unique linked regulators, and high-confidence i-modulons (precision > 0.5, recall > 0.25). (b) The precision of ICA-derived i-modulons are higher than sparse-PCA-derived i-modulons for both datasets (p -value < .01, Mann-Whitney U test). Boxplot whiskers represent extrema of data, box bounds represent upper and lower quartiles, and center-line represents the median value. (c) Scatterplot of CysB i-modulon activities against Cbl+CysB i-modulon activities. The Cbl+CysB i-modulon was only active under sulfur starvation conditions. (d) Distribution of gene regulation in the combined NagC/TyrR i-modulon. (e) Distribution of gene regulation in the combined GntR/TyrR i-modulon. (f) Distribution of tryptophan-related regulatory mechanisms in the TrpR i-modulon. Gene names in blue indicate positive gene coefficients, and gene names in red indicate negative gene coefficients. (g) Distribution of leucine- and isoleucine-related regulatory mechanisms in the Leu/Ile i-modulon. High concentrations of leucine or isoleucine either charge tRNAs that pause transcription to form termination loops, or activate the transcription factor IlvY to repress transcription of *ilvC*. Subscripts indicate pseudogene fragments.

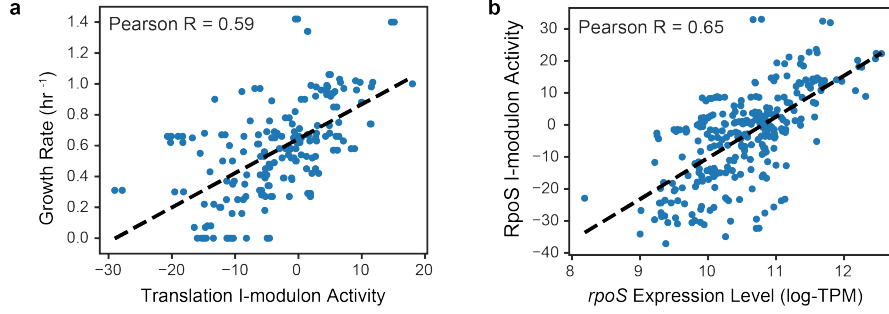


Figure A.5: Correlations related to the fear-greed trade-off. (a) The Translation i-modulon activity level was correlated with growth rate. (b) The RpoS i-modulon activity level was correlated with *rpoS* gene expression.

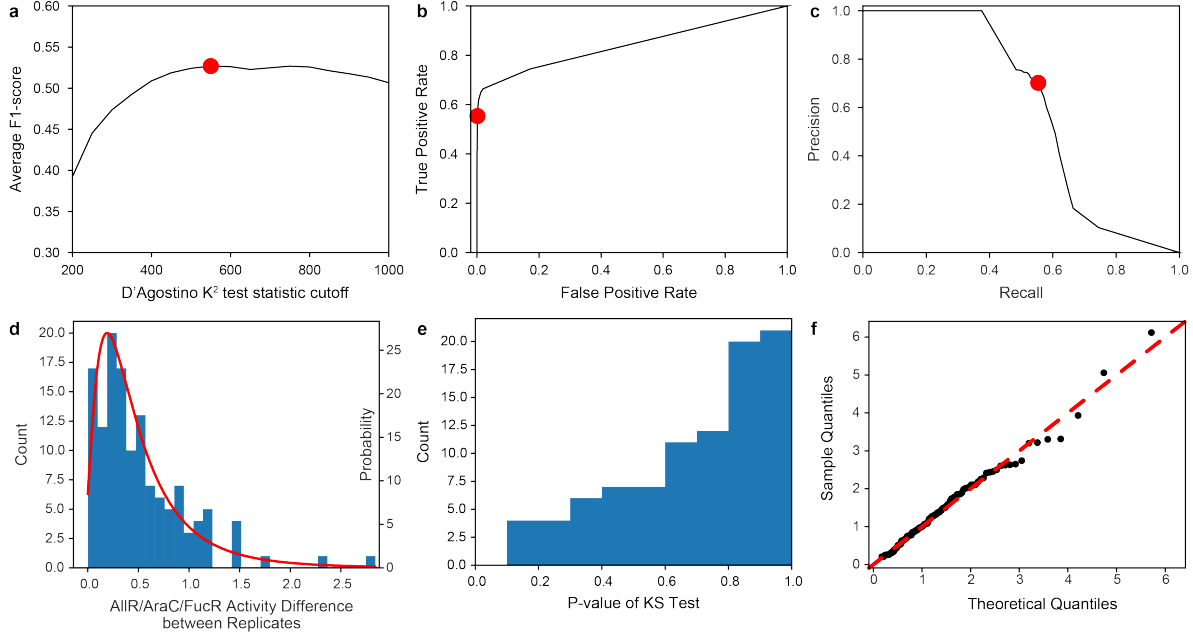


Figure A.6: Significance thresholds for components and differential activity. (a) Sensitivity analysis on the D'Agostino K^2 test statistic cutoff. Red dot indicates optimal value. (b) Precision-recall curve and (c) receiver operating characteristic (ROC) curve from varying the cutoff. (d) The difference in ArgR i-modulon activities between biological replicates follows a log-normal distribution (shown in red). (e) P-values from K-S tests on each i-modulon are higher than 0.1, indicating that an inability to reject the null hypothesis that the activity differences between replicates are from a log-normal distribution. (f) Quantile-quantile plot of difference in AllR/AraC/FucR i-modulon activities between biological replicates provides further evidence of log-normal distribution.

EvgA I-Modulon

Regulated by: EvgA

Biological Function: Acid and osmotic stress response

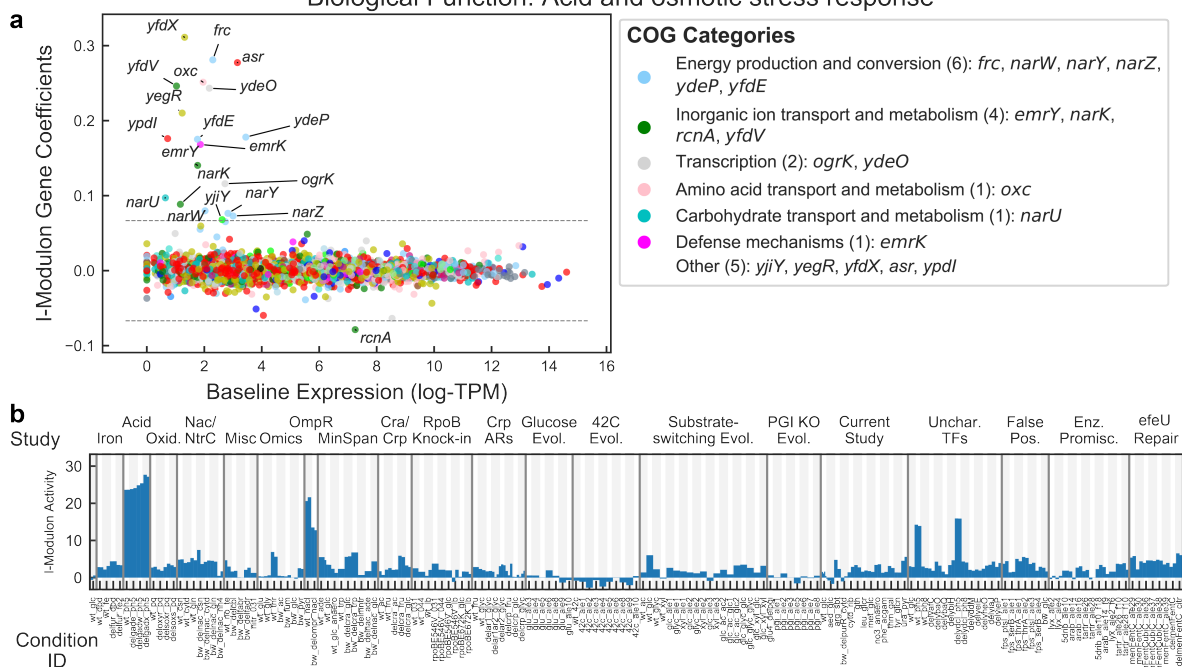


Figure A.7: Descriptive characteristics of the EvgA i-modulon with explanations. Similar characteristics are shown for all 92 i-modulons in Chapter B. (a) Scatterplot of i-modulon gene coefficients against gene expression (log-TPM) in the reference condition (*E. coli* MG1655 grown on glucose M9 minimal medium). Genes are colored by their cluster of orthologous groups (COG) categories. (b) I-modulon activities across the entire PRECISE compendium, grouped by original study. Each condition occupies a constant width, regardless of the number of biological replicates.

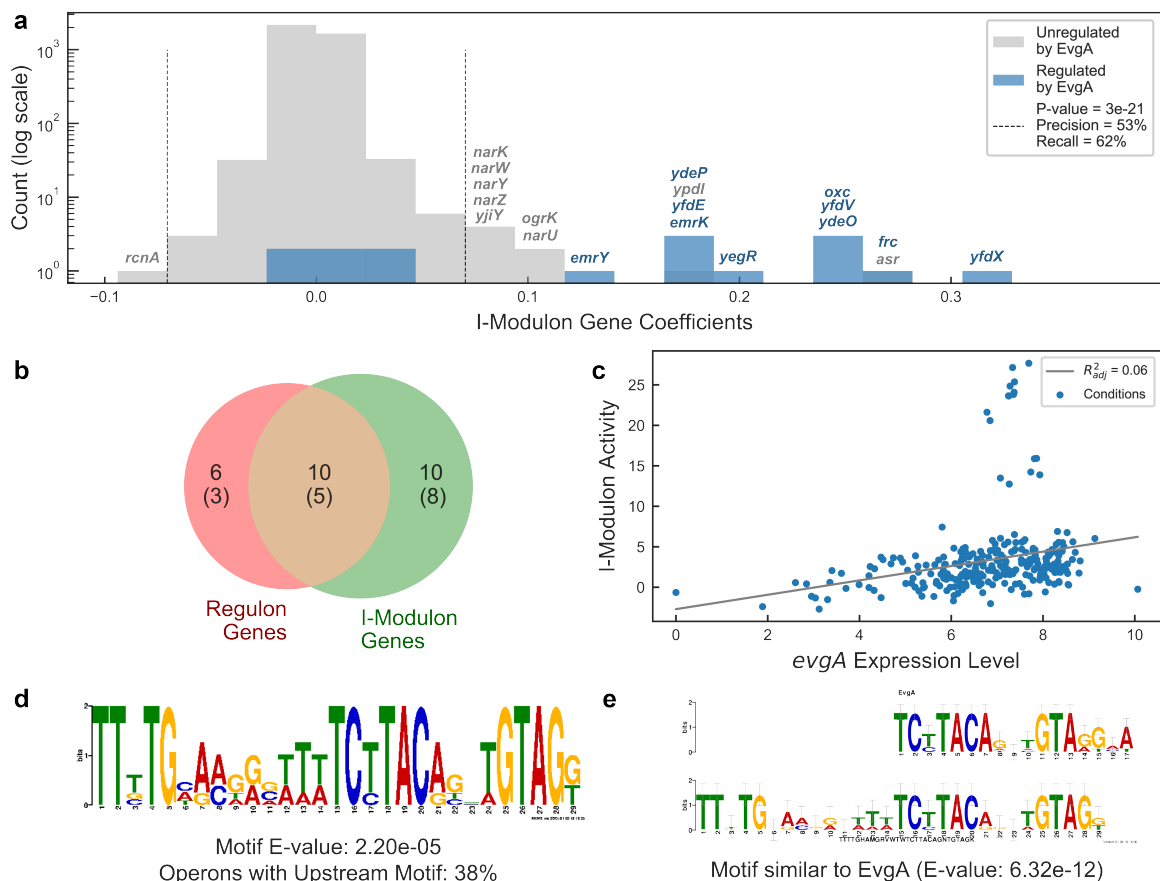


Figure A.8: Continuation of Figure A.7(a) Histogram of i-modulon gene coefficients on a logarithmic count scale. Two histograms are overlaid, one for genes with reported regulation by the associated regulator, and one for genes without reported regulation. If multiple regulators are associated, additional histograms are overlaid. Gene names are listed above bars outside of the significance threshold (dashed vertical line). Statistics on regulator enrichment are reported in the legend. (b) Venn diagram of genes in associated regulon vs. significant genes in i-modulon. Numbers in parenthesis indicate number of operons in each category. (c) Scatterplot of i-modulon activity against expression level of associated regulator. Regulator knock-outs, if applicable, are indicated in orange. The R^2_{adj} accounts for minimum expression levels, as discussed in Figure A.5c and Methods. (d) Motif identified upstream of i-modulon genes, if found. (e) Similar motif in RegulonDB to the motif identified in panel d, if found.

A.2 References

1. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. & Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. en. *PLoS Biol.* **5**, e8 (Jan. 2007).
2. Baneyx, F. & Mujacic, M. Recombinant protein folding and misfolding in Escherichia coli. en. *Nat. Biotechnol.* **22**, 1399–1408 (Nov. 2004).

Appendix B

Descriptive characteristics of the 92 i-modulons extracted from PRECISE

This appendix contains detailed information on each i-modulon. Each page characterizes one i-modulon, which are sorted by name. The EvgA i-modulon page is described in Figures A.7 and A.8. EvgA is known to be activated under acid and osmotic stress, which is consistent with its i-modulon activities.

B.1 Automated characterization of i-modulons

The scatter plot (Figure A.7a) compares the i-modulon gene weights to the reference expression profile (defined as the average log-TPM of sample IDs *control_wt_glc_1* and *control_wt_glc_2*). The horizontal dashed lines represent the threshold separating significant and

non-significant genes in an i-modulon. Each gene is colored by its functional category, defined by clusters of orthologous groups of genes queried from EggNOG[1]. Gene names are provided on the scatter plot if the total number of significant genes is below 25. Subscripts indicate split genes in *E. coli* MG1655, classified as pseudogenes.

The bar chart (Figure A.7b) displays the i-modulon activities across all conditions in the database, grouped by study. Each condition shares the same total width, regardless of the number of biological replicates.

The histograms (Figure A.8a) show the gene weights in an i-modulon on a log-scale, split by enriched regulon. The vertical bars represent the significance threshold for the component. The gray histogram captures genes that are not in an enriched regulon. The brown histogram captures genes that are in more than one enriched regulon.

The venn diagram (Figure A.8b) compares genes in an i-modulon (above the threshold) to genes in the enriched regulons. If the i-modulon name contains a slash (/), the regulon circle contains the union of genes in any enriched regulon (i.e. genes in either regulon). If the i-modulon name contains a plus sign (+), the regulon circle contains the intersection of the regulons (i.e. genes in both regulons). Numbers in parentheses indicate number of operons (complete or partial) in each category. Operons were defined using RegulonDB. This panel is hidden if no enriched regulons were identified.

The regulon scatter plot (Figure A.8c) compares the expression level of enriched regulators to the i-modulon activity. Each point represents an expression profile. Two lines were fit to each scatter plot, a simple line and a broken line. The broken line represents a minimal expression level required before a correlation is observed between the i-modulon activity and the regulator

expression. Only the line with the highest R_{adj}^2 is shown, where

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

The broken line is modeled as:

$$y = \begin{cases} ac + b & \text{if } x < c \\ ax + b & \text{if } x \geq c \end{cases}$$

where x is the expression level of the gene encoding the TF, and y is the i-modulon activity level.

Parameters a , b , and c are optimized using the *optimize.curve_fit* function in the Python *scipy* package[2]. Three initial values for c were tested to identify the optimal fit: minimum expression, maximum expression, and mean expression.

This panel is hidden if no enriched regulons were identified, or if expression levels were not measured for the regulator (see Thiamine i-modulon). If there are two or more enriched regulons with expression levels, the two TFs with the strongest association to the i-modulon are shown. Experimental conditions with knocked-out TFs are marked in orange and were not used to find the best-fit line.

B.1.1 Motif Search

Motifs for (Figure A.8d) were identified by searching upstream sequences using MEME[3]. Genes in each i-modulon were grouped into operons. A stringent upstream sequence was defined as the region from -300 to the start of the operon. We searched for zero or one motif per sequence using an E-value threshold of 10^{-3} , searching for motifs with all widths between 6 and 30 basepairs, and minimum number of found motifs equal to 1/3 the total number of operons for all non-genomic i-modulons. For i-modulons with no enriched motifs in the stringent upstream sequence, we searched for motifs in a broader upstream sequence of -600 to +100, keeping all

other parameters constant. The E-value of the motif and the percent of i-modulon operons with the upstream motif are listed below the motif.

The motif comparison (Figure A.8e) was generated by comparing i-modulon motifs against known motifs from RegulonDB using TOMTOM[4], with an E-value threshold of 0.01 and allowing incomplete matches. The E-value of the comparison is shown below the figure.

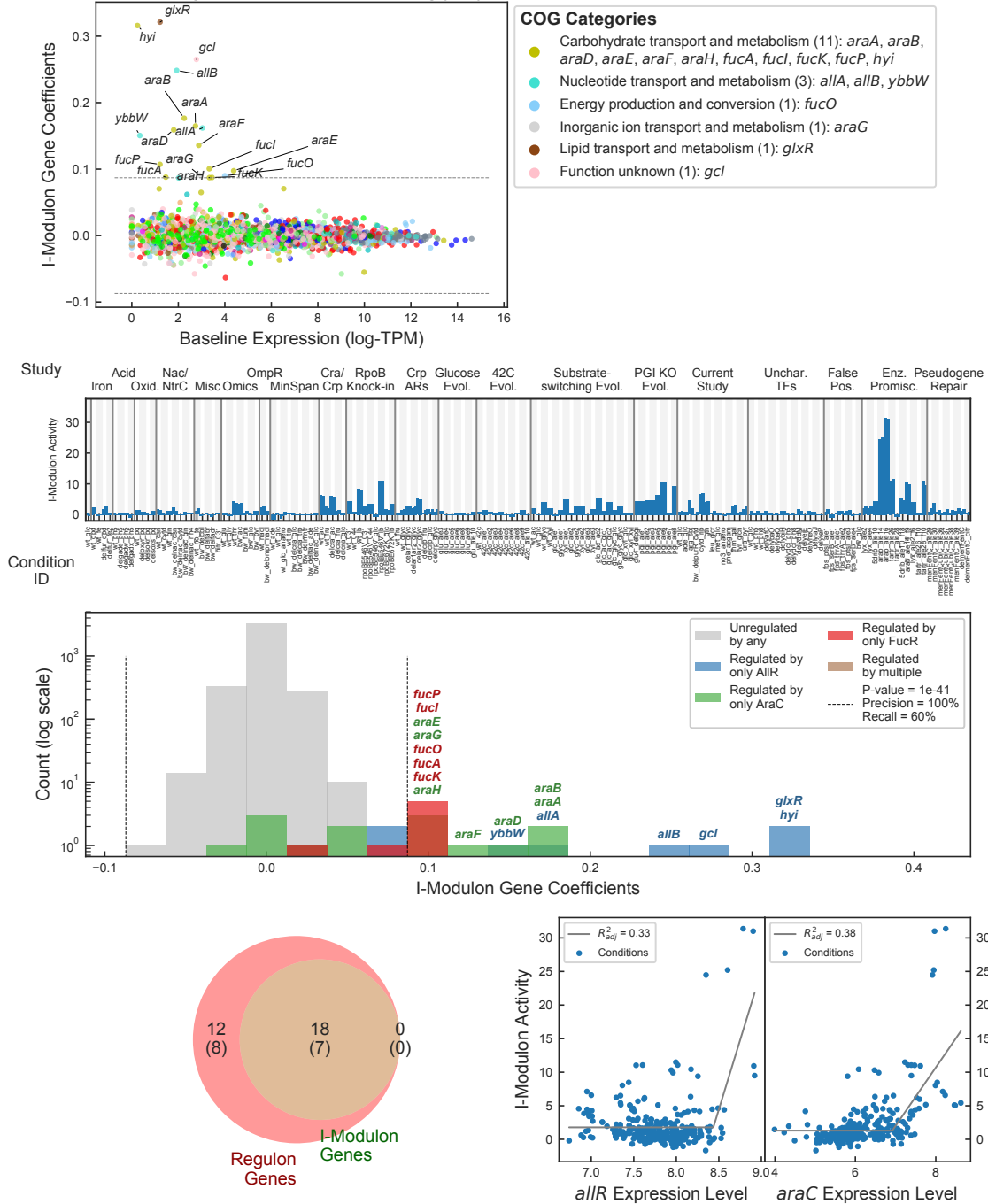
B.2 References

1. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. en. *Nucleic Acids Res.* **44**, D286–93 (Jan. 2016).
2. Oliphant, T. E. Python for Scientific Computing. *Computing in Science Engineering* **9**, 10–20 (May 2007).
3. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. en. *Nucleic Acids Res.* **37**, W202–8 (July 2009).
4. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. en. *Genome Biol.* **8**, R24 (2007).

AlIR/AraC/FucR I-Modulon

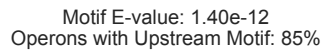
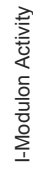
Regulated by: AlIR or AraC or FucR

Biological Function: Allantoin, glyoxylate, L-arabinose, and L-fucose catabolism



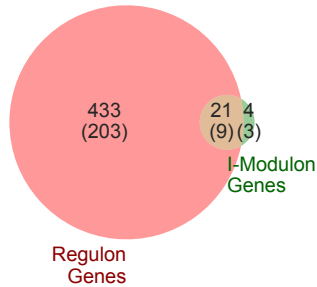
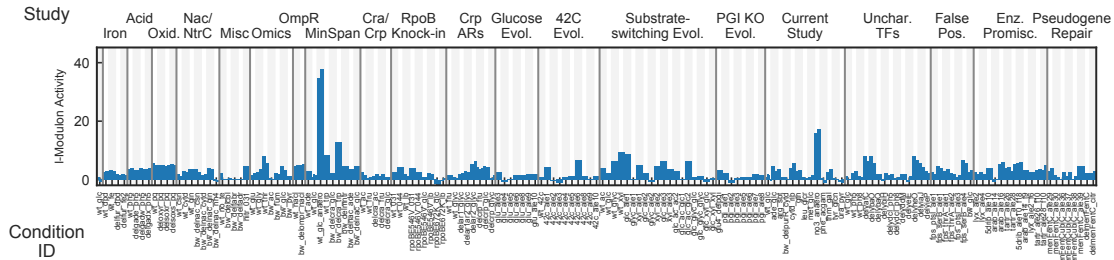
Regulated by: ArcA
Biological Function: Anaerobic response

Biological Function: Anaerobic response



Regulated by: ArcA
Biological Function: Electron Transport Chain

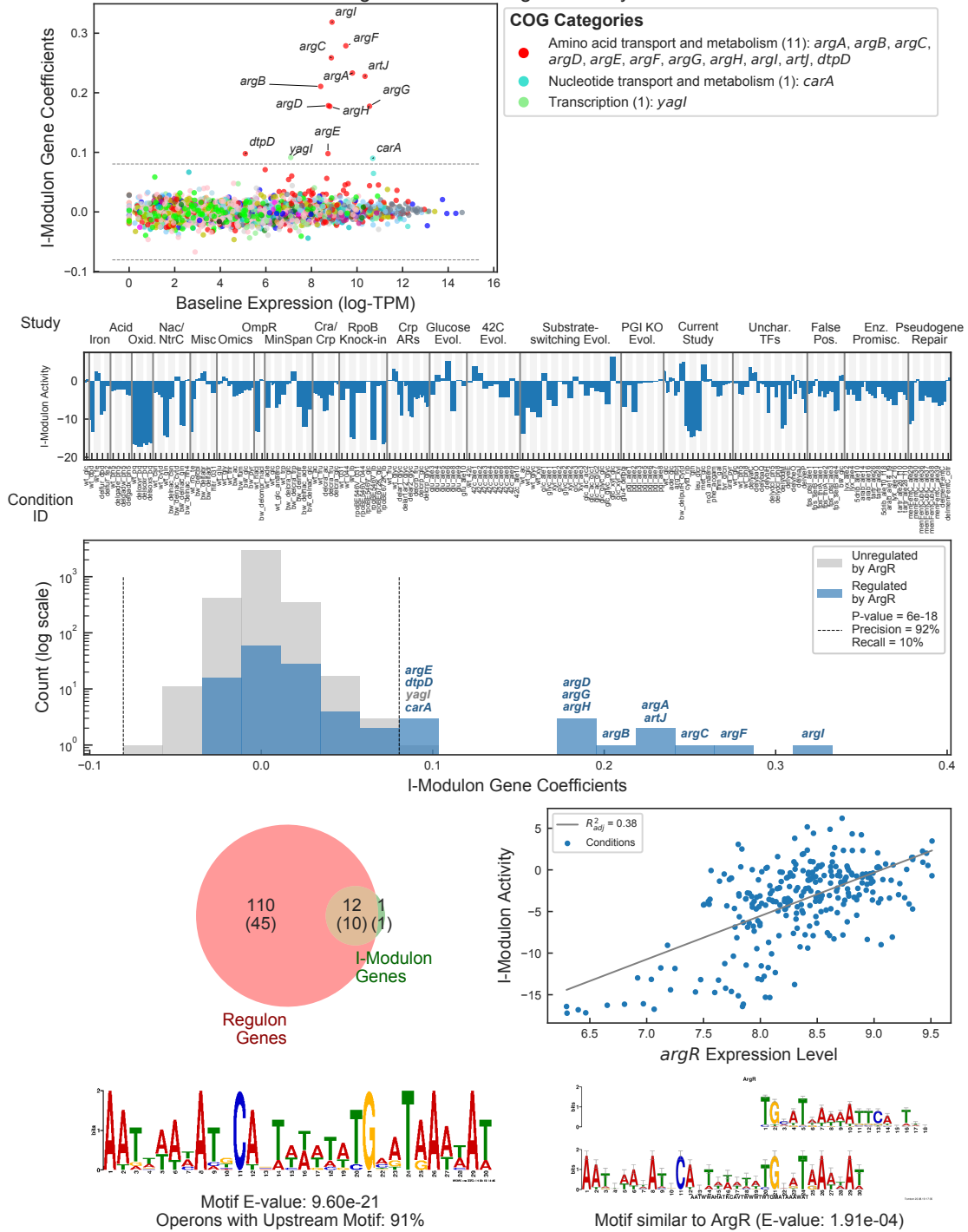
Biological Function: Electron Transport Chain



ArgR I-Modulon

Regulated by: ArgR

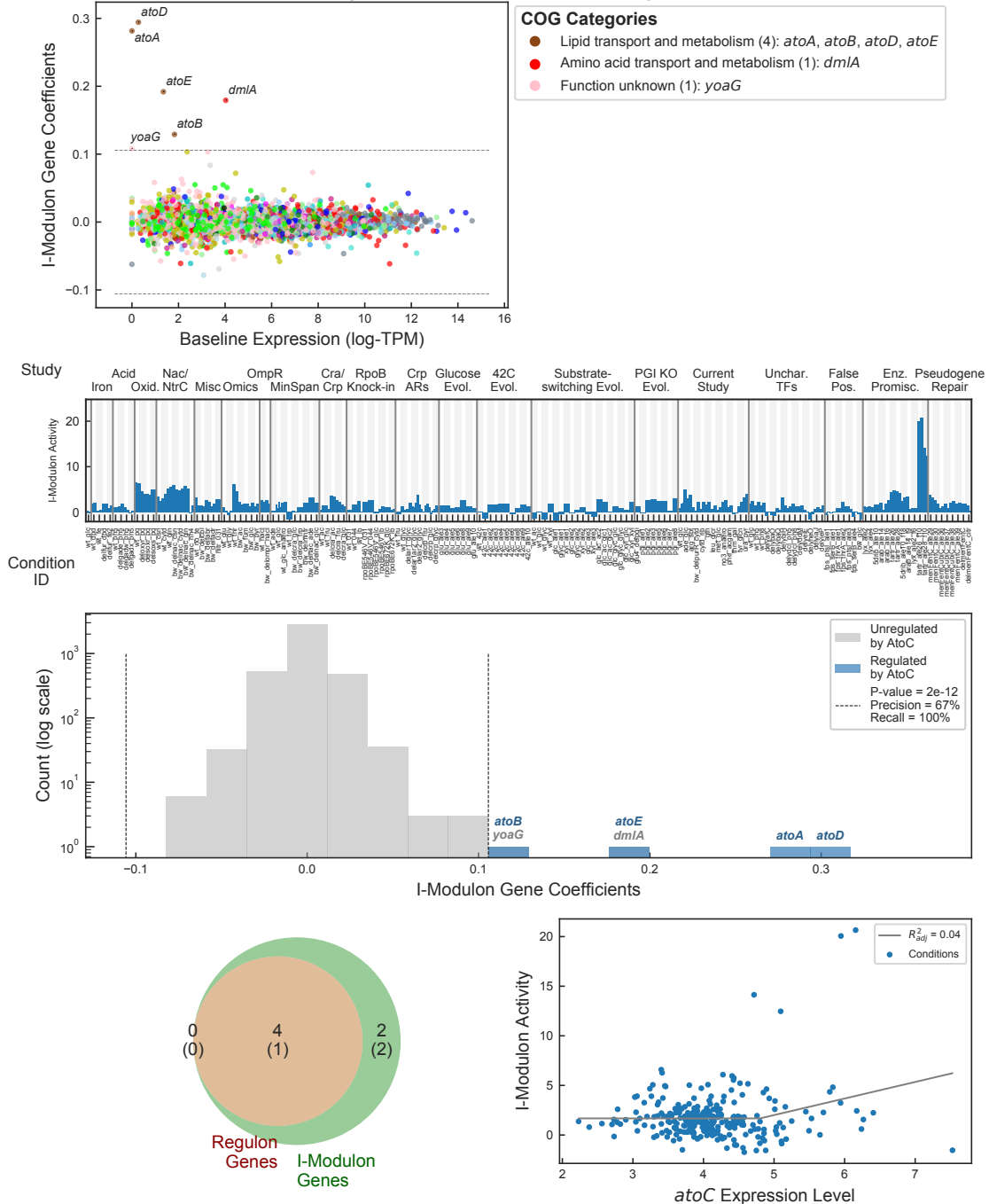
Biological Function: Arginine biosynthesis



AtoC I-Modulon

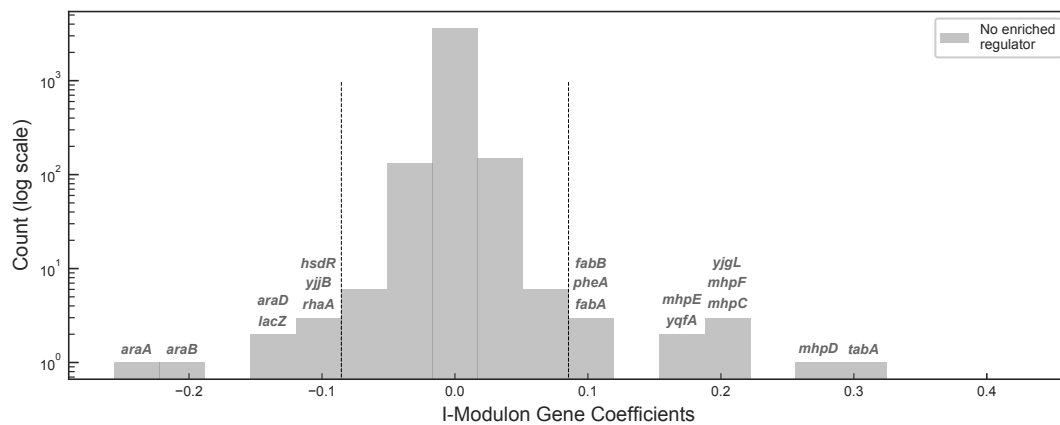
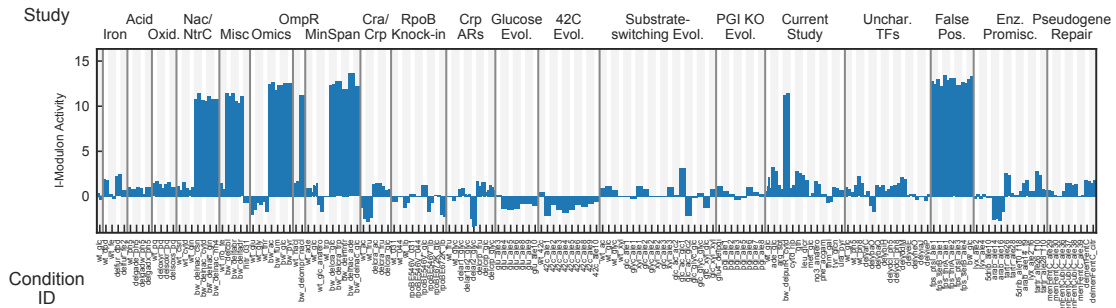
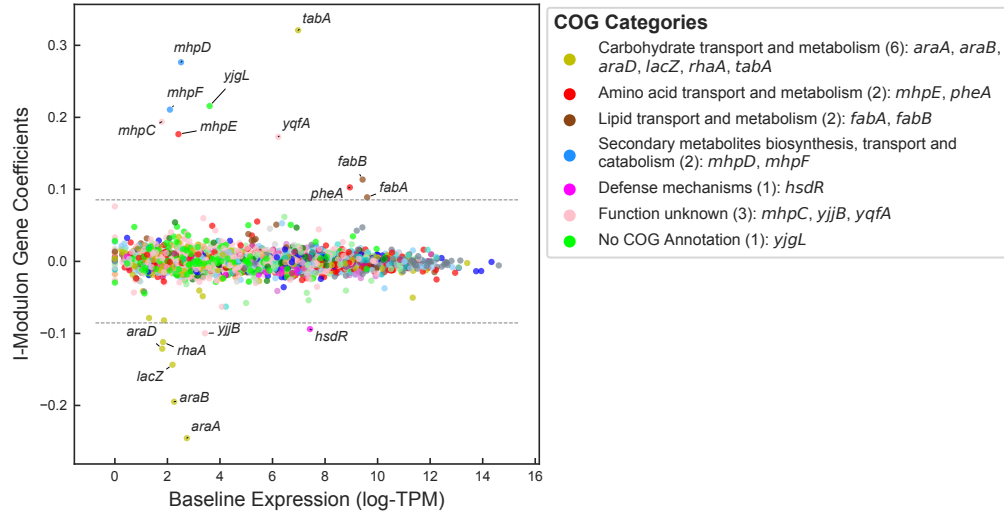
Regulated by: AtoC

Biological Function: Acetoacetate degradation



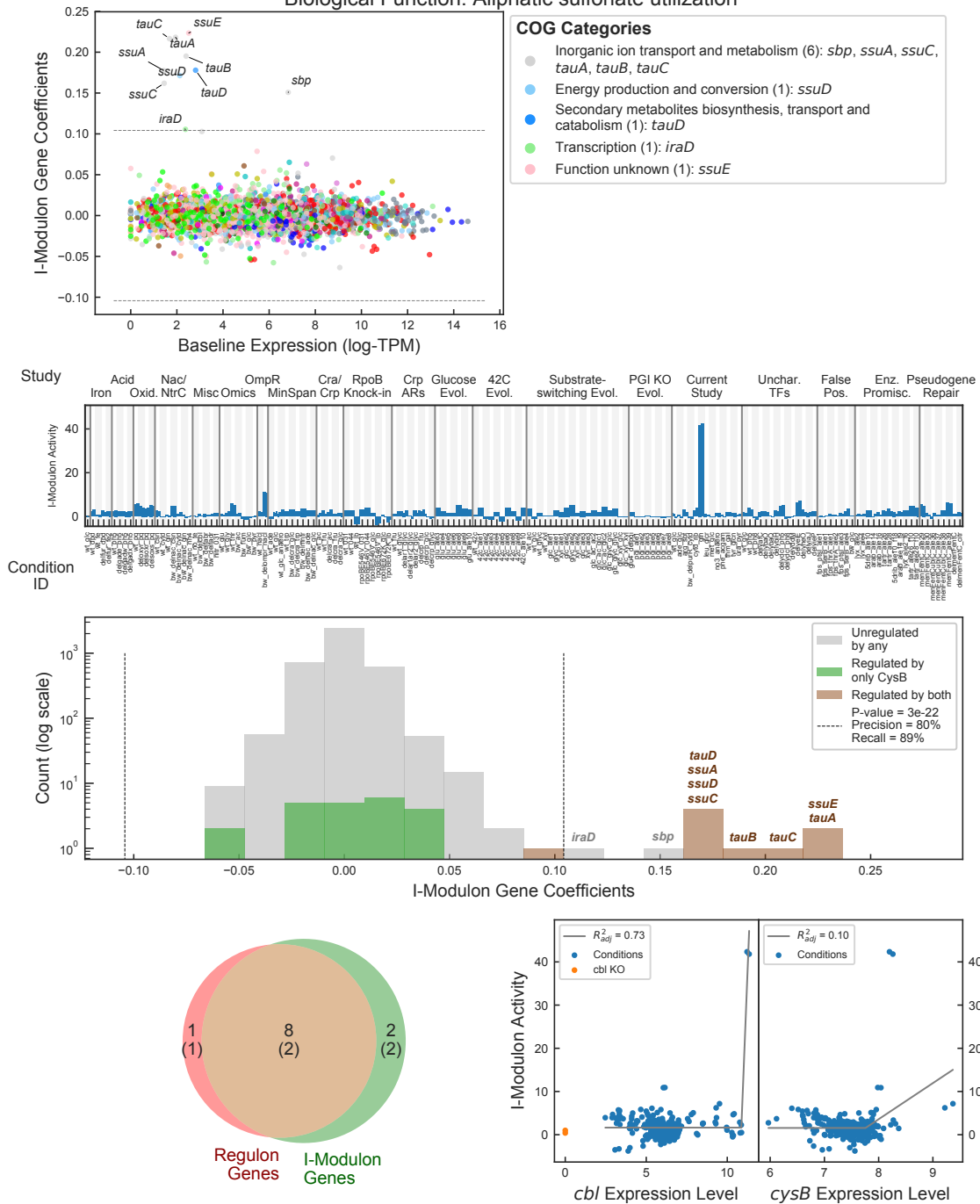
BW25113 I-Modulon

Biological Function: Transcriptional difference between BW25113 and MG1655



Cbl + CysB I-Modulon

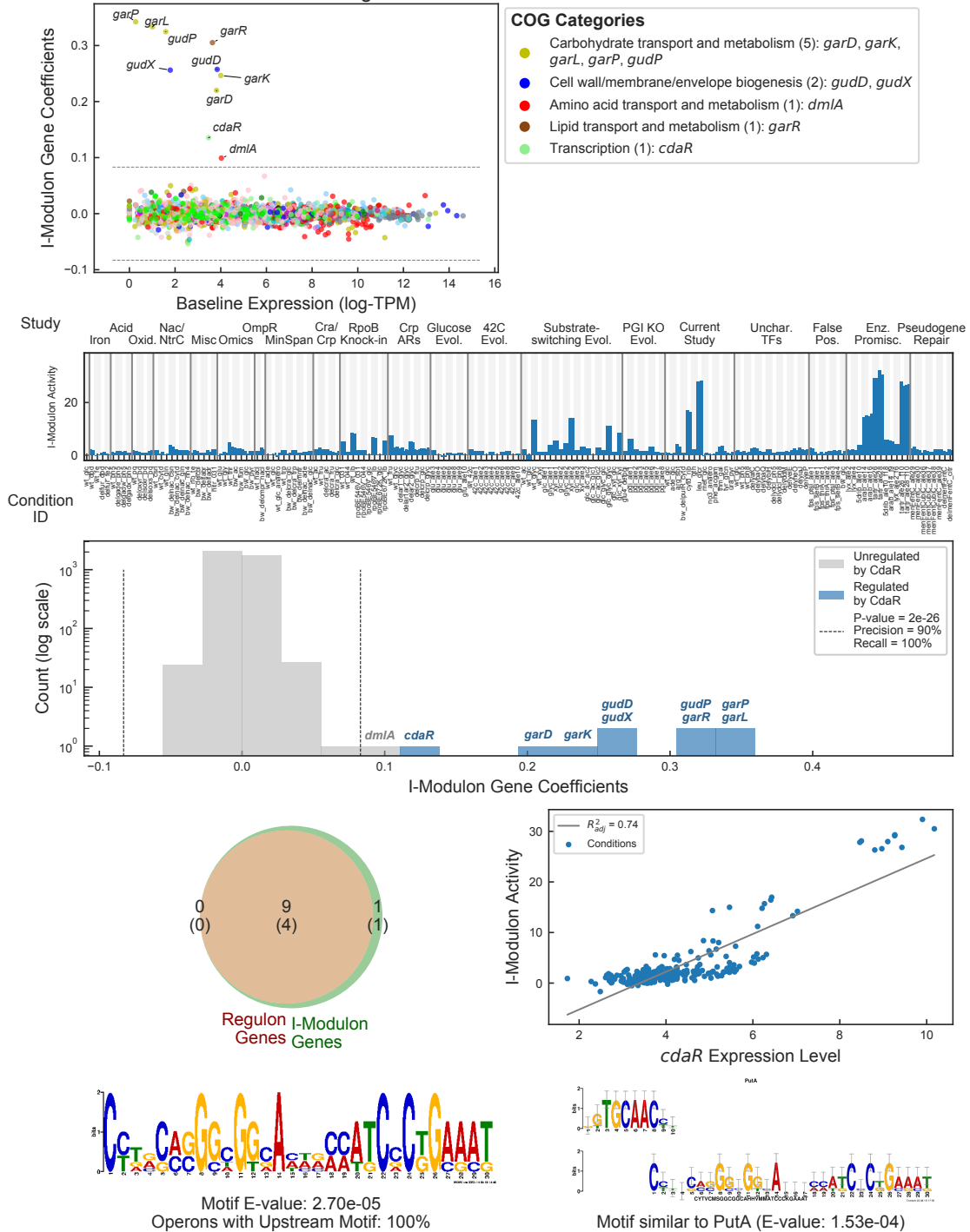
Regulated by: Cbl and CysB
Biological Function: Aliphatic sulfonate utilization



CdaR I-Modulon

Regulated by: CdaR

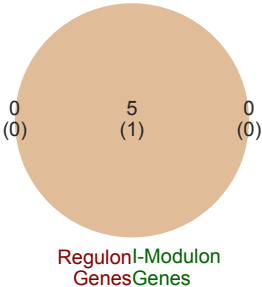
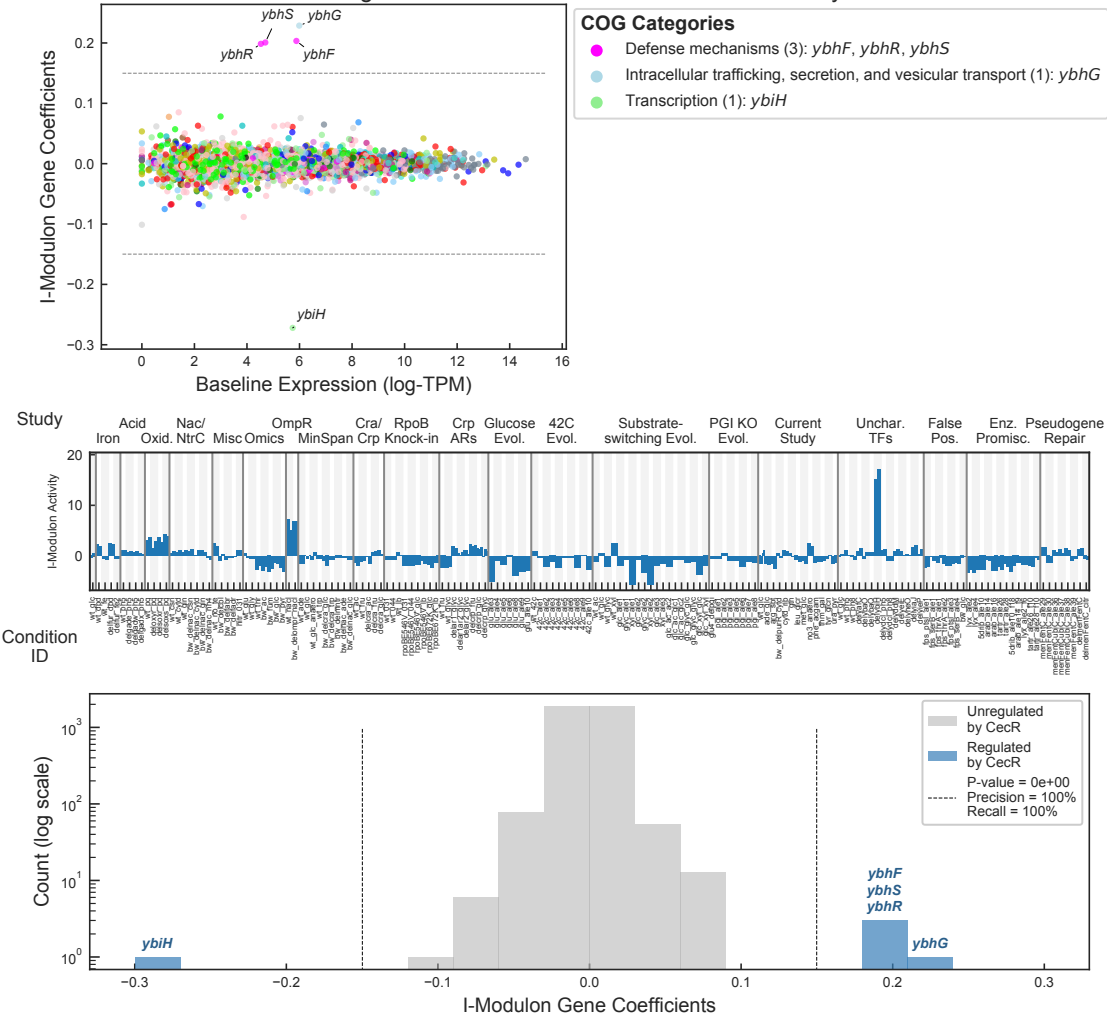
Biological Function: Glucarate catabolism



CecR I-Modulon

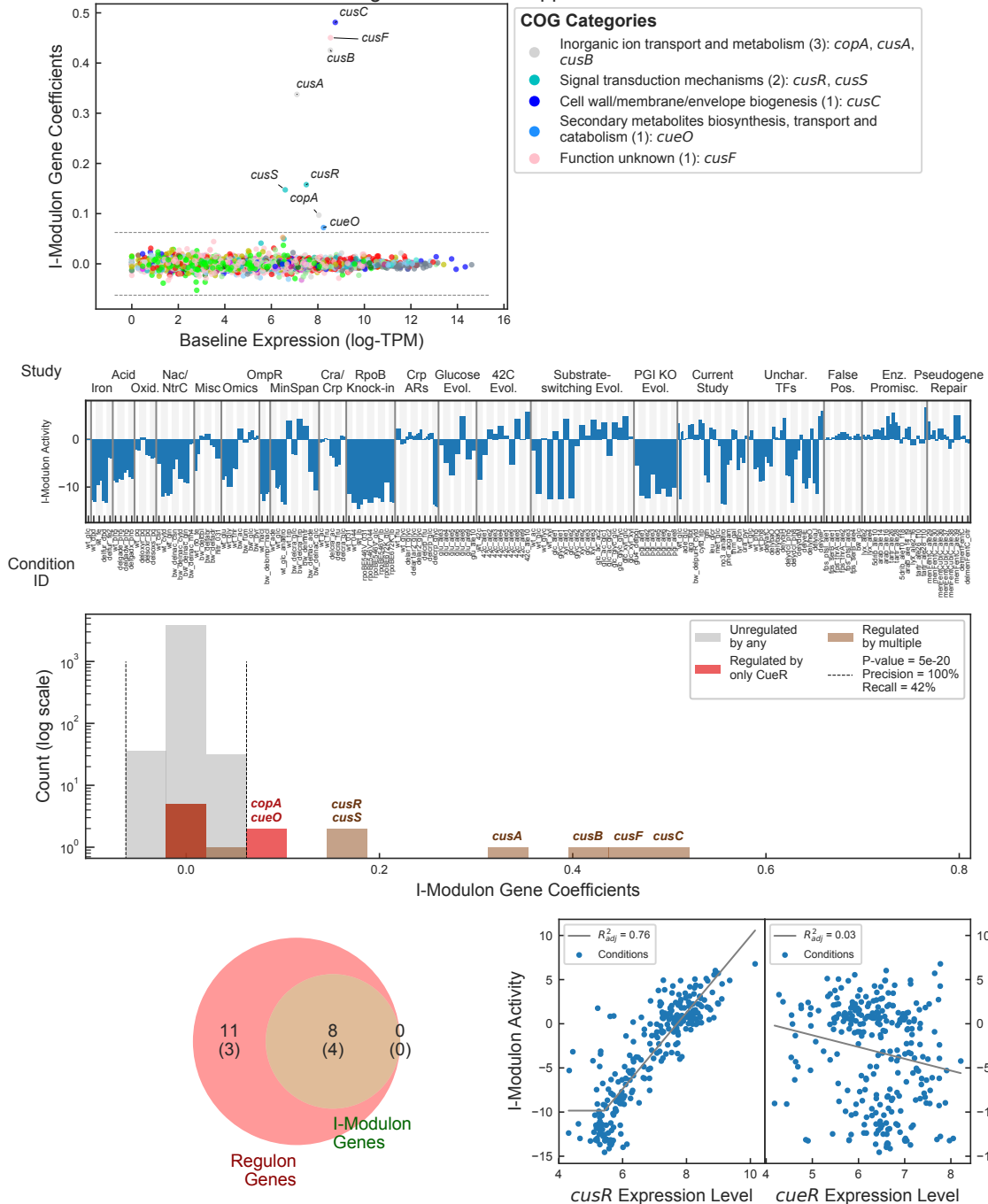
Regulated by: CecR

Biological Function: Related to antibiotic sensitivity



Copper I-Modulon

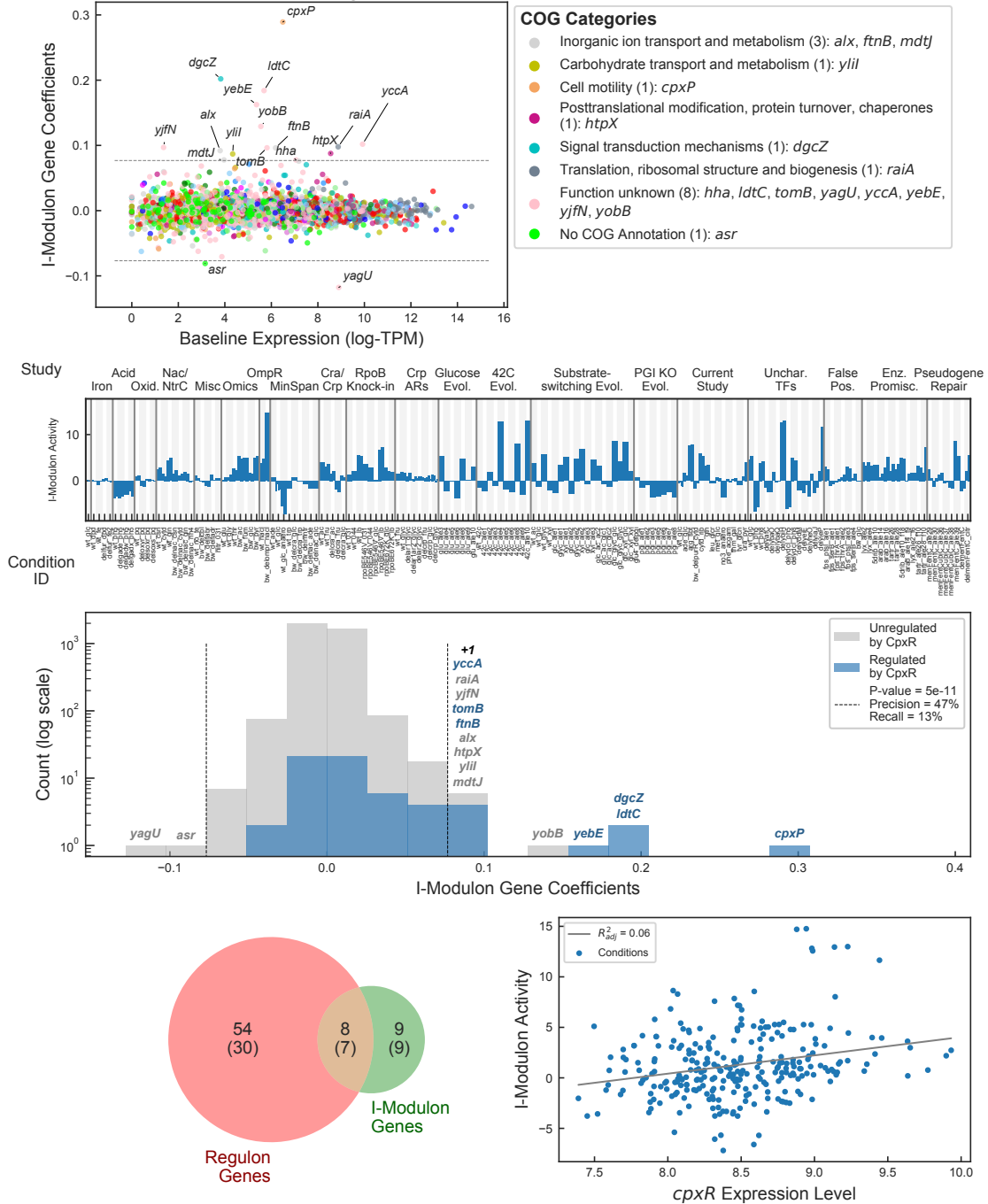
Regulated by: CusR or HprR or CueR
Biological Function: Copper homeostasis



CpxR I-Modulon

Regulated by: CpxR

Biological Function: Various stress responses



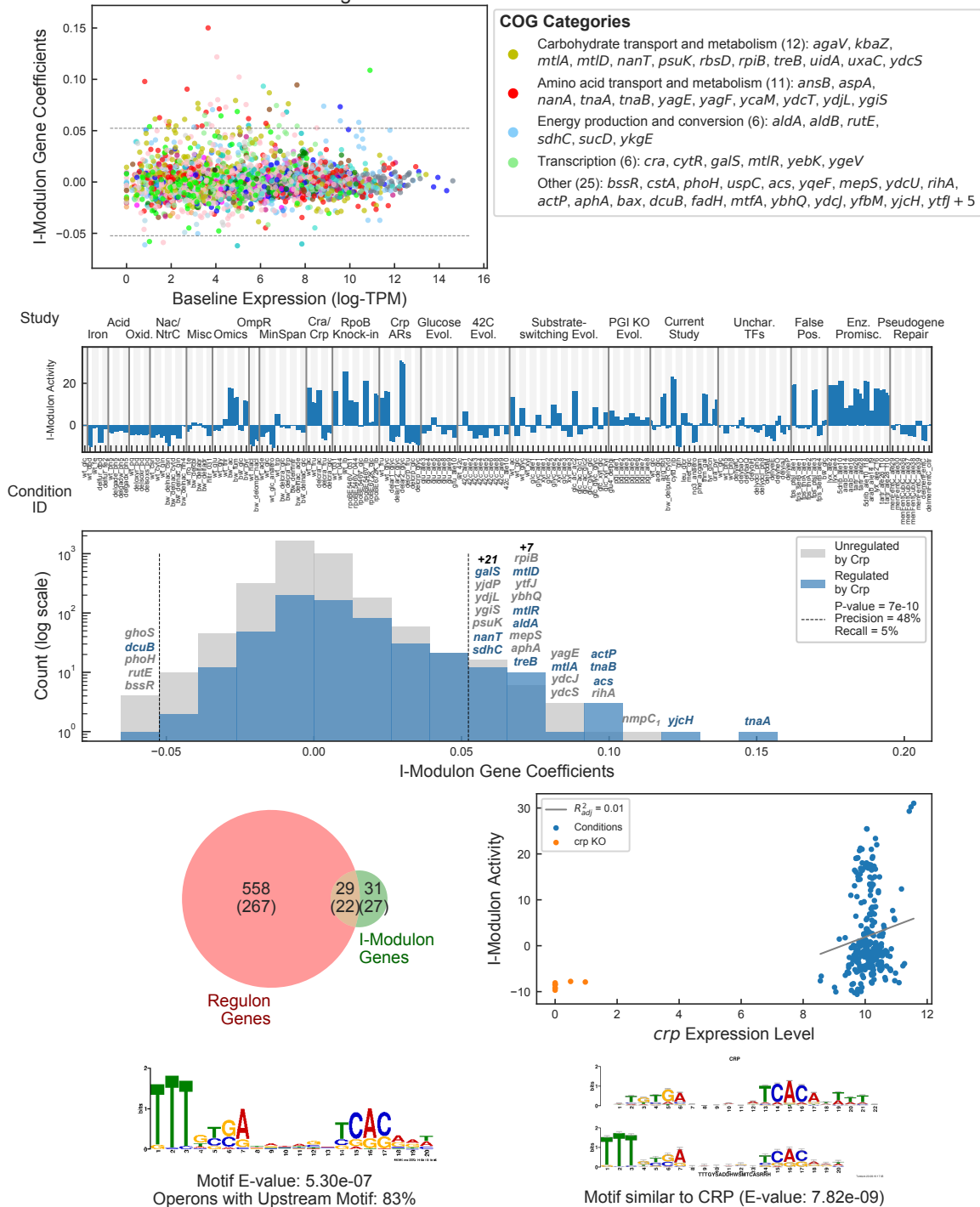
Regulated by: Cra
Biological Function: Central carbon metabolism



Crp – 1 I-Modulon

Regulated by: Crp

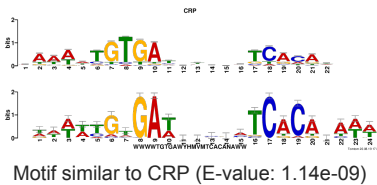
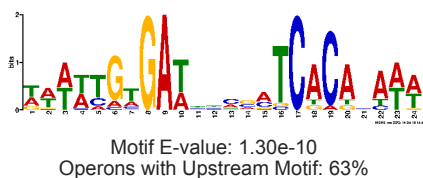
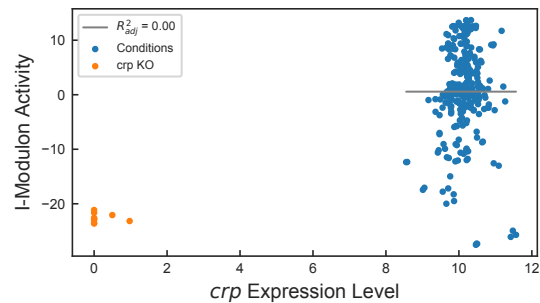
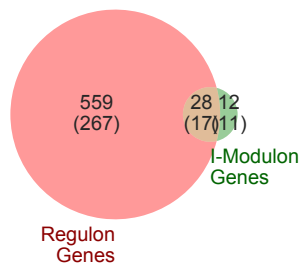
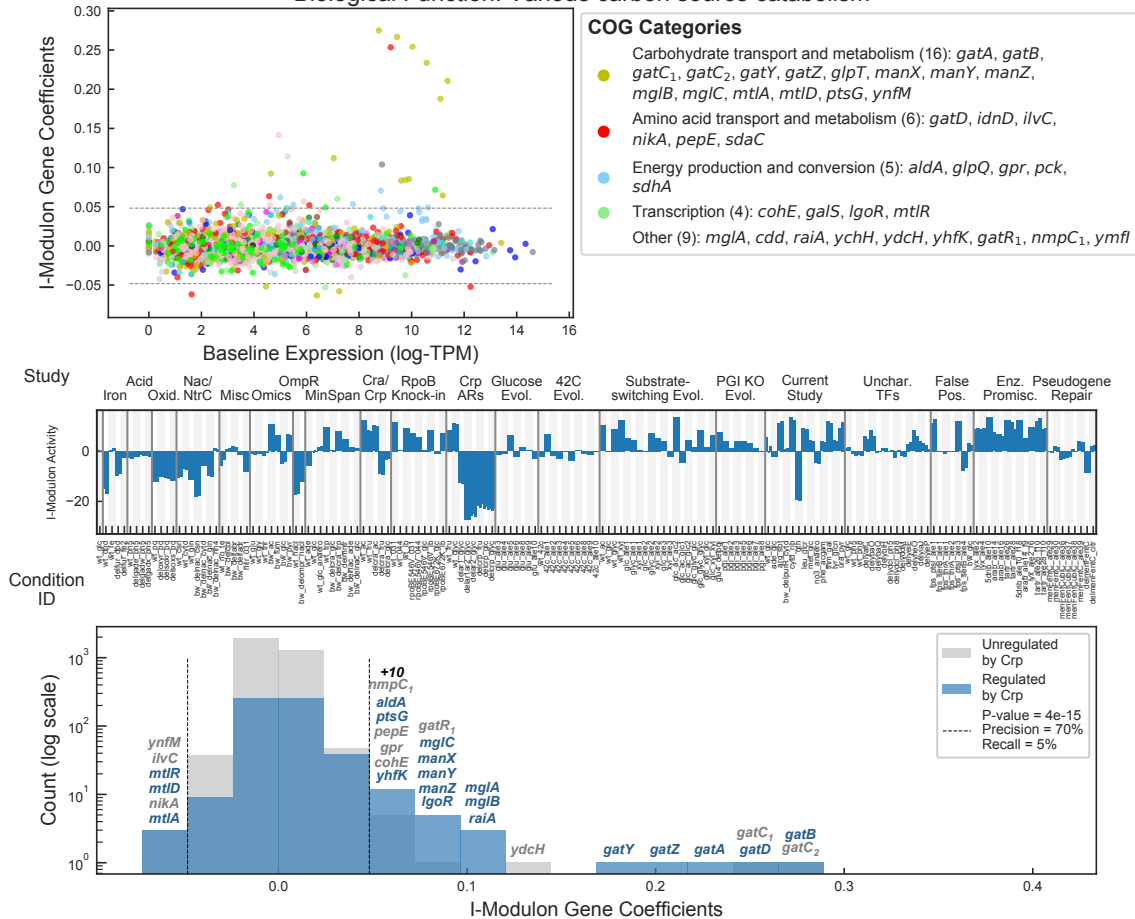
Biological Function: Miscellaneous functions



Crp – 2 I-Modulon

Regulated by: Crp

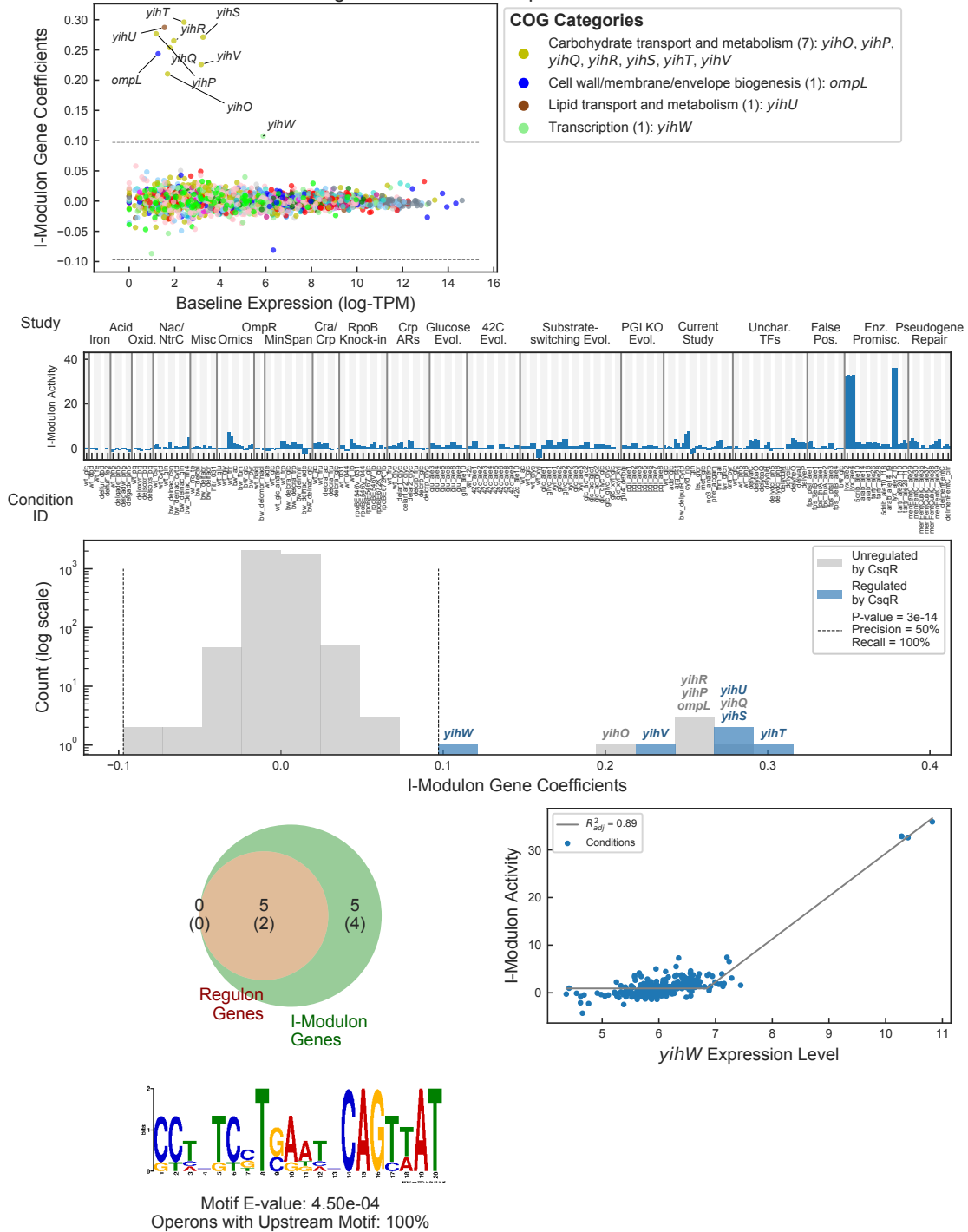
Biological Function: Various carbon source catabolism



CsqR I-Modulon

Regulated by: CsqR

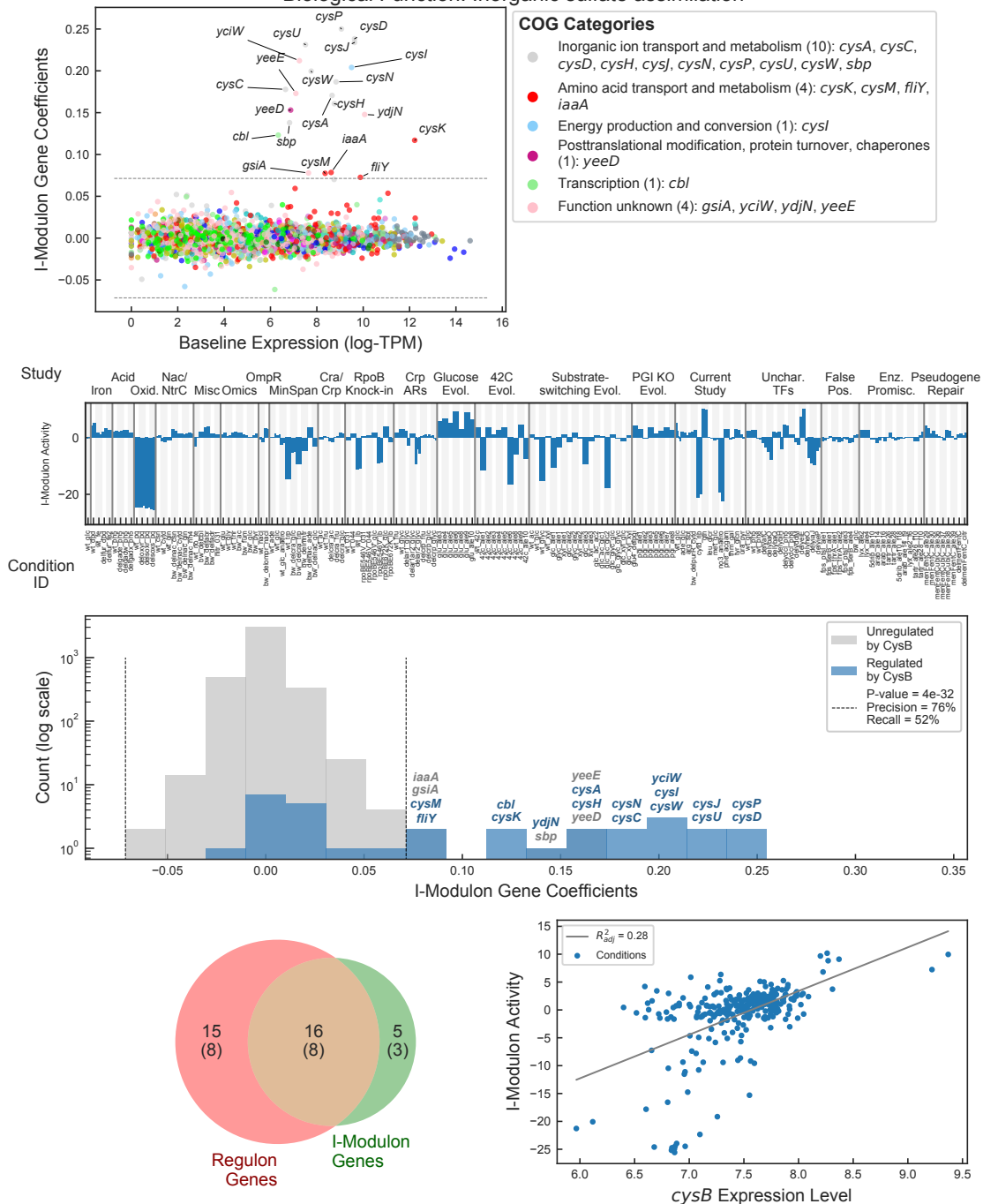
Biological Function: Sulfoquinovose catabolism



CysB I-Modulon

Regulated by: CysB

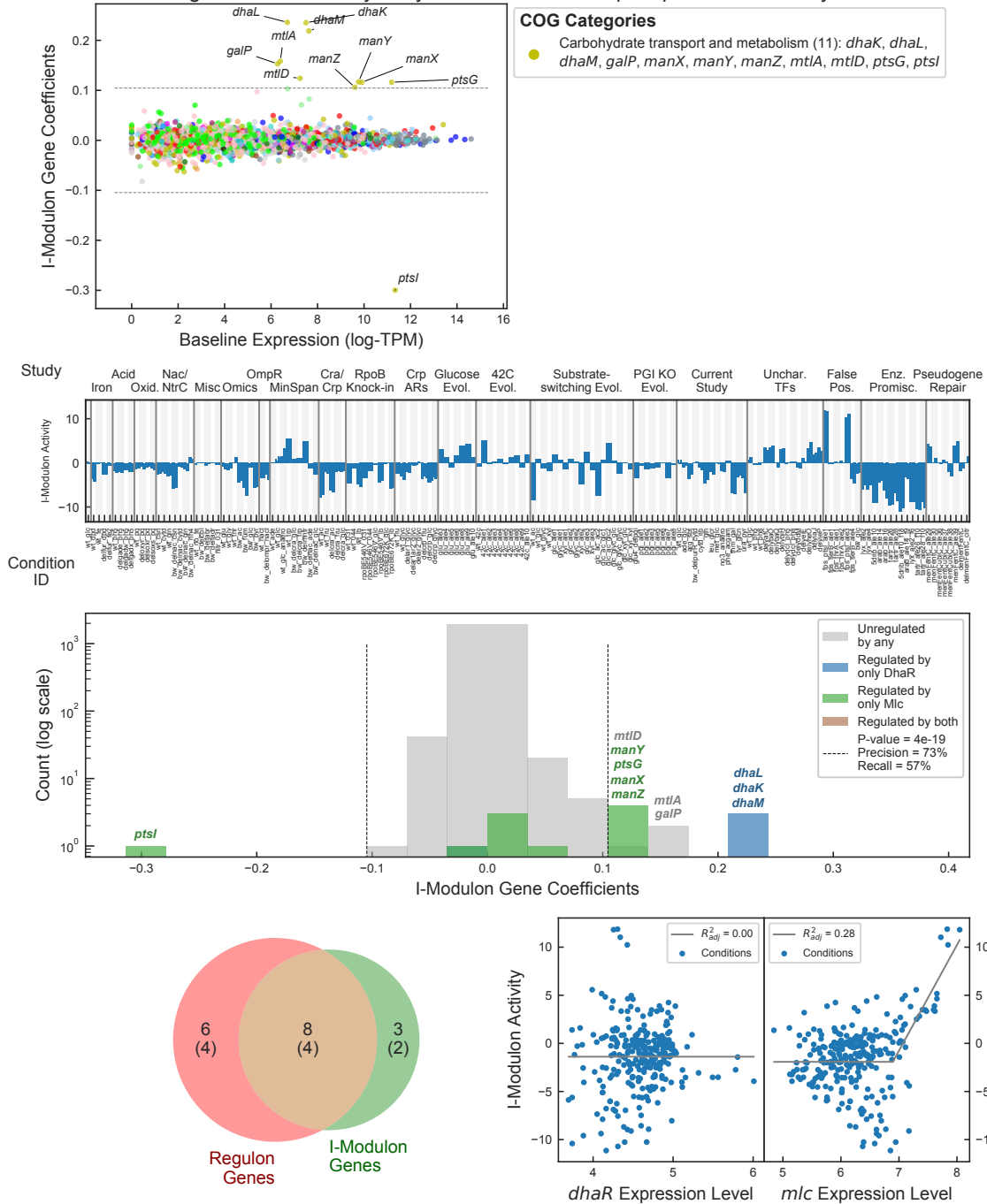
Biological Function: Inorganic sulfate assimilation



DhaR/Mlc I-Modulon

Regulated by: DhaR or Mlc

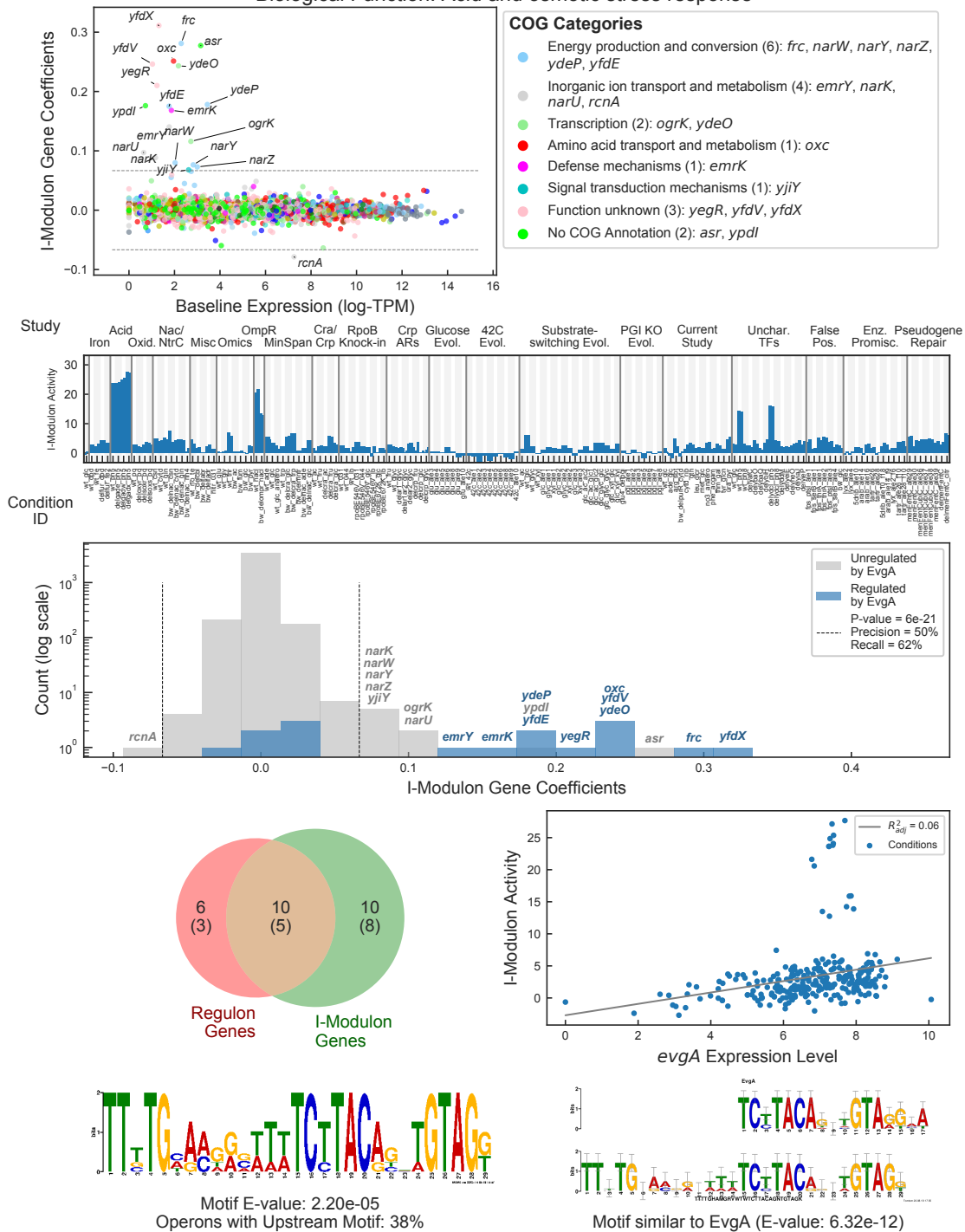
Biological Function: Dihydroxyacetone kinase and phosphotransferase systems



EvgA I-Modulon

Regulated by: EvgA

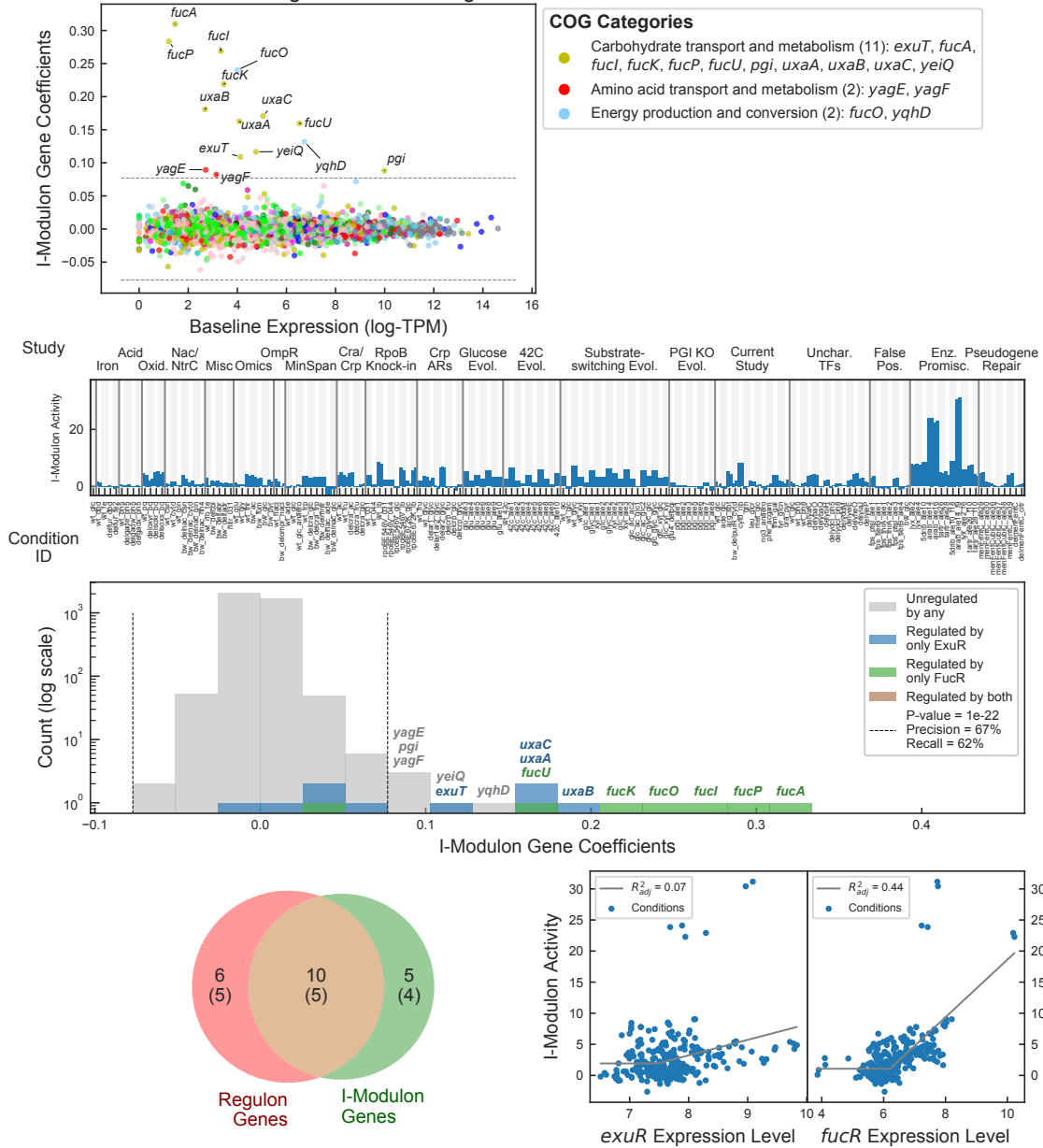
Biological Function: Acid and osmotic stress response



ExuR/FucR I-Modulon

Regulated by: ExuR or FucR

Biological Function: D-galacturonate and L-fucose catabolism

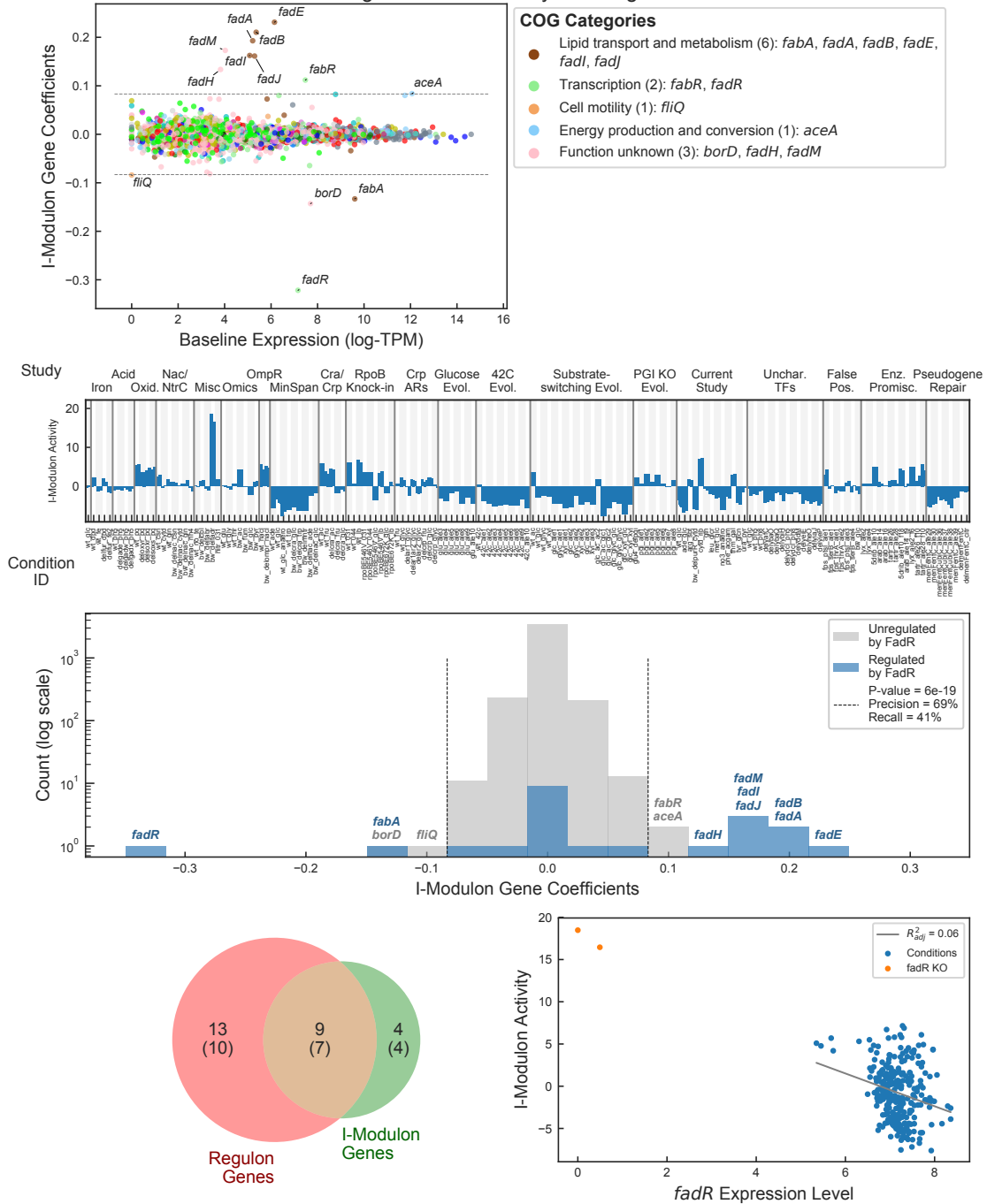


Operons with Upstream Motif: 67%

FadR I-Modulon

Regulated by: FadR or IclR

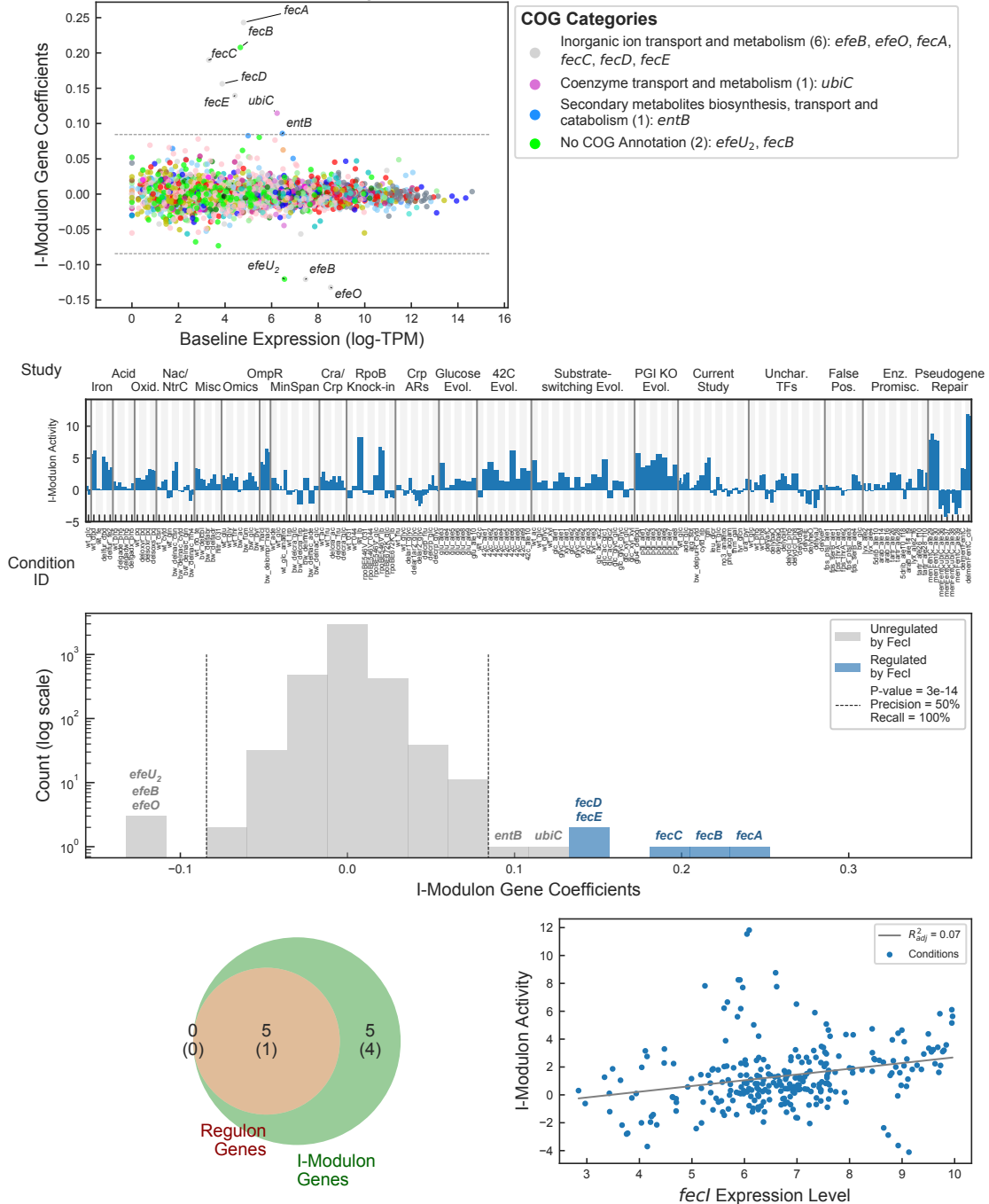
Biological Function: Fatty acid degradation



FecI I-Modulon

Regulated by: FecI

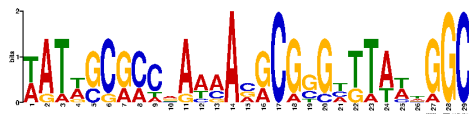
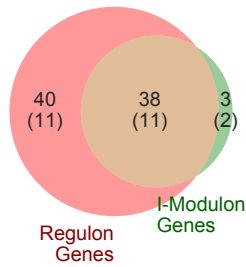
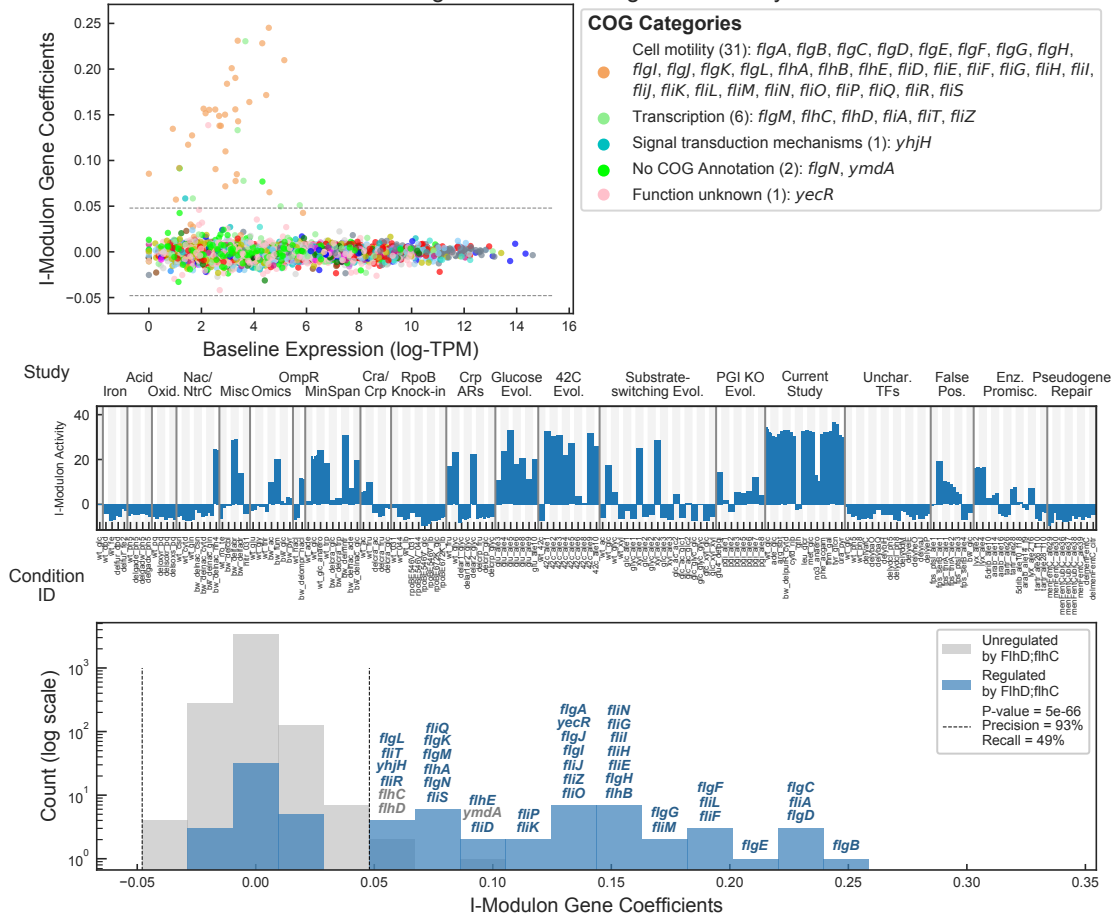
Biological Function: Ferric citrate transport



FlhDC I-Modulon

Regulated by: FlhDC

Biological Function: Flagella assembly

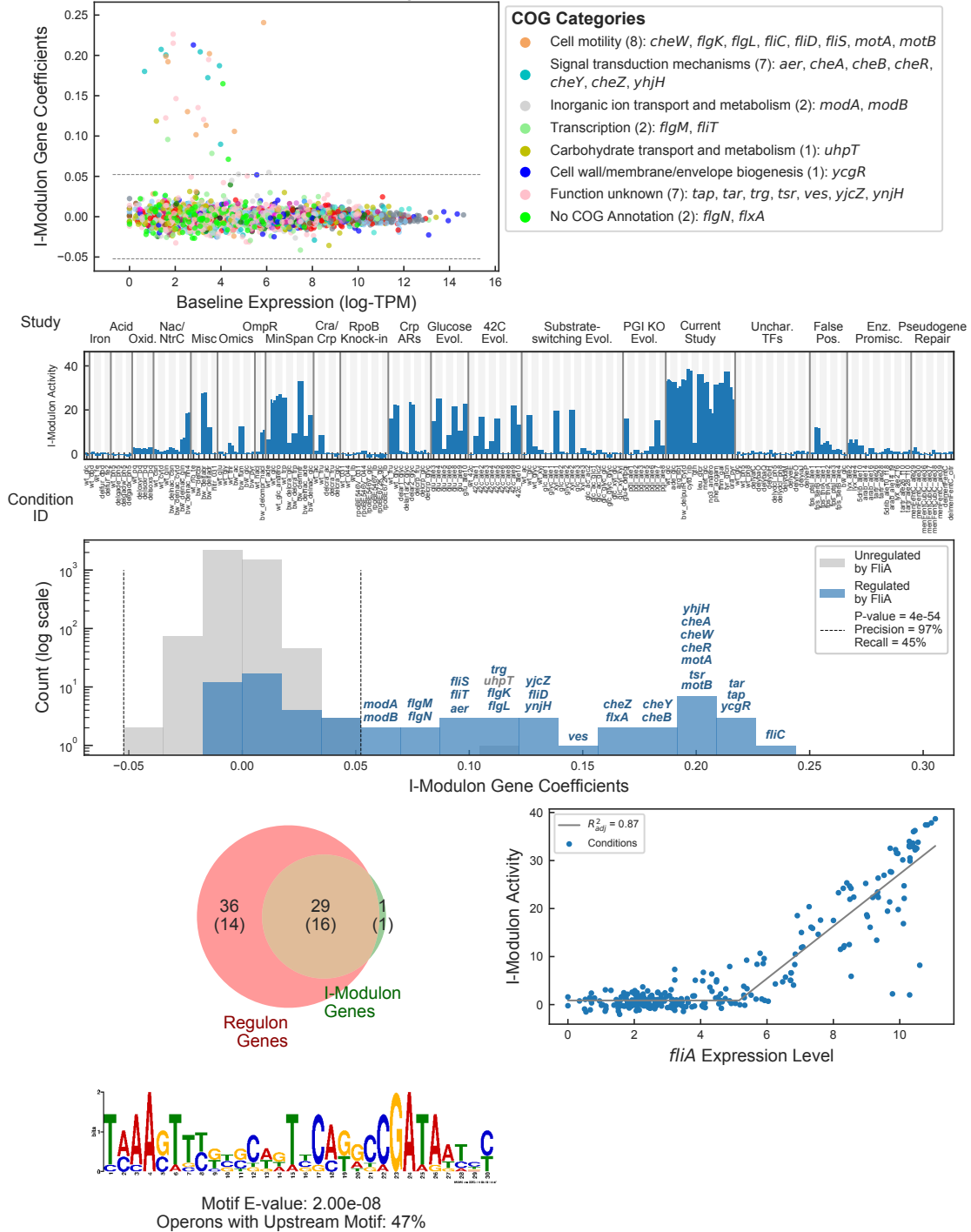


Motif E-value: 7.40e-07
Operons with Upstream Motif: 54%

FliA I-Modulon

Regulated by: FliA

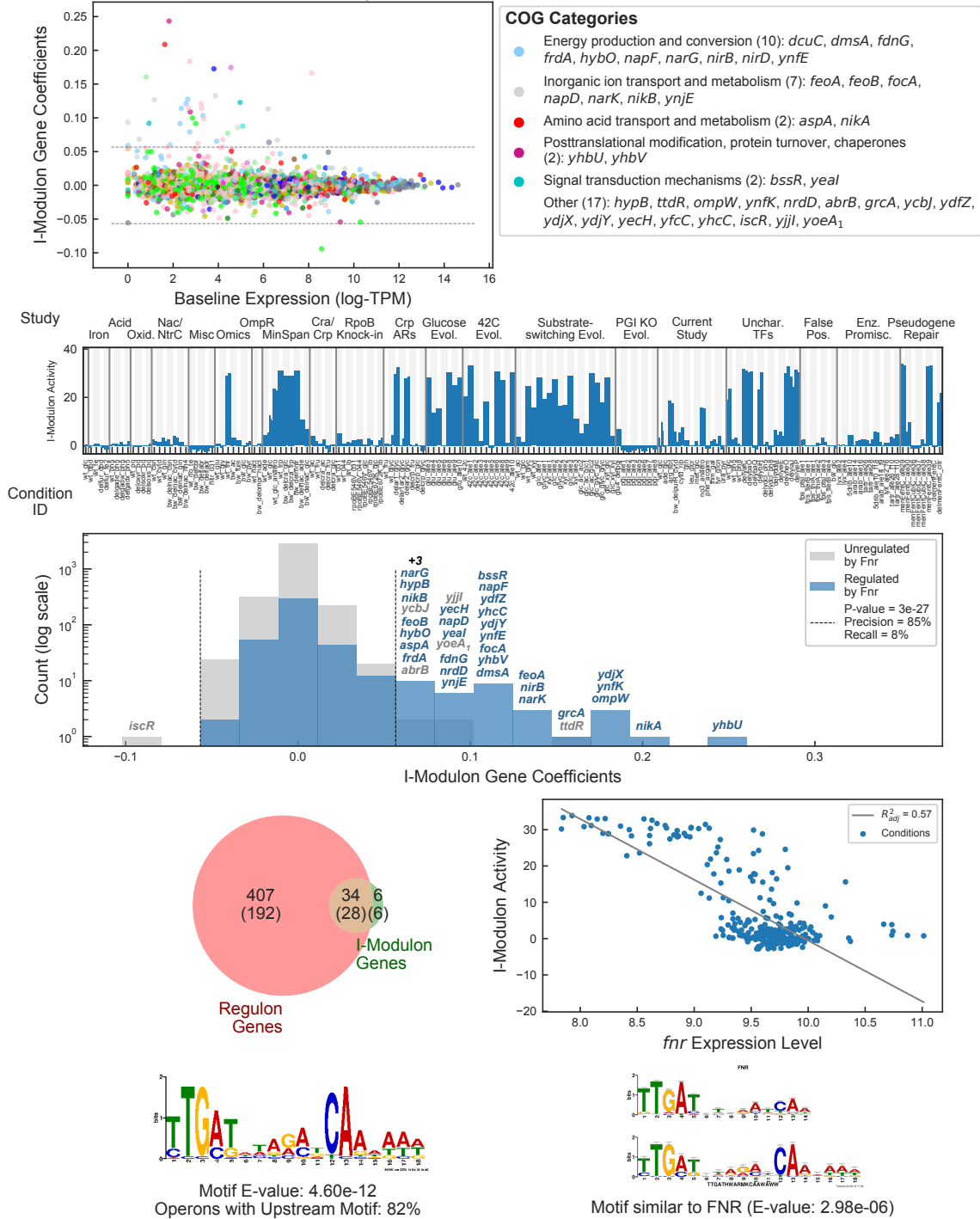
Biological Function: Chemotaxis



Fnr I-Modulon

Regulated by: Fnr

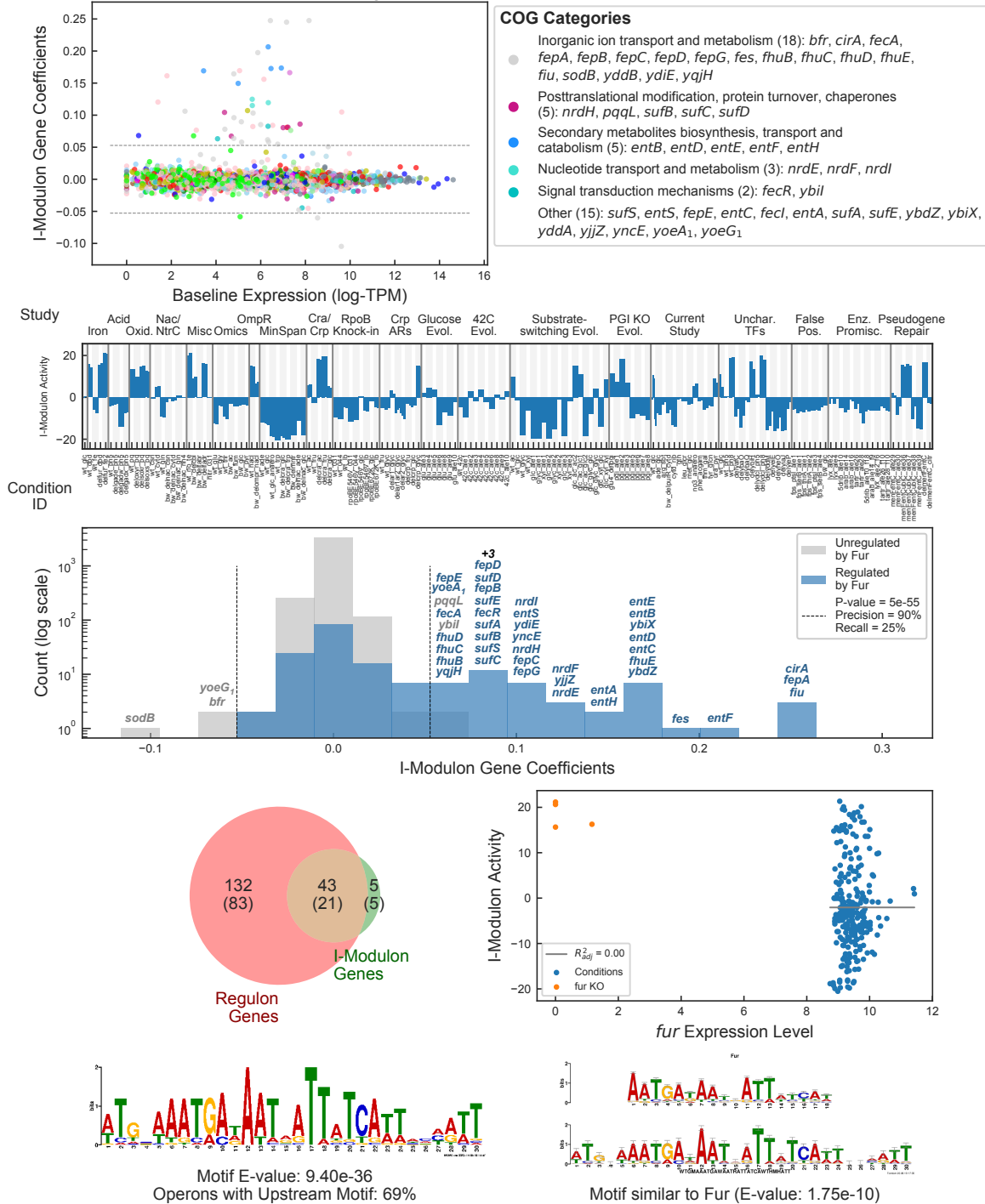
Biological Function: Anaerobic response



Fur – 1 I-Modulon

Regulated by: Fur

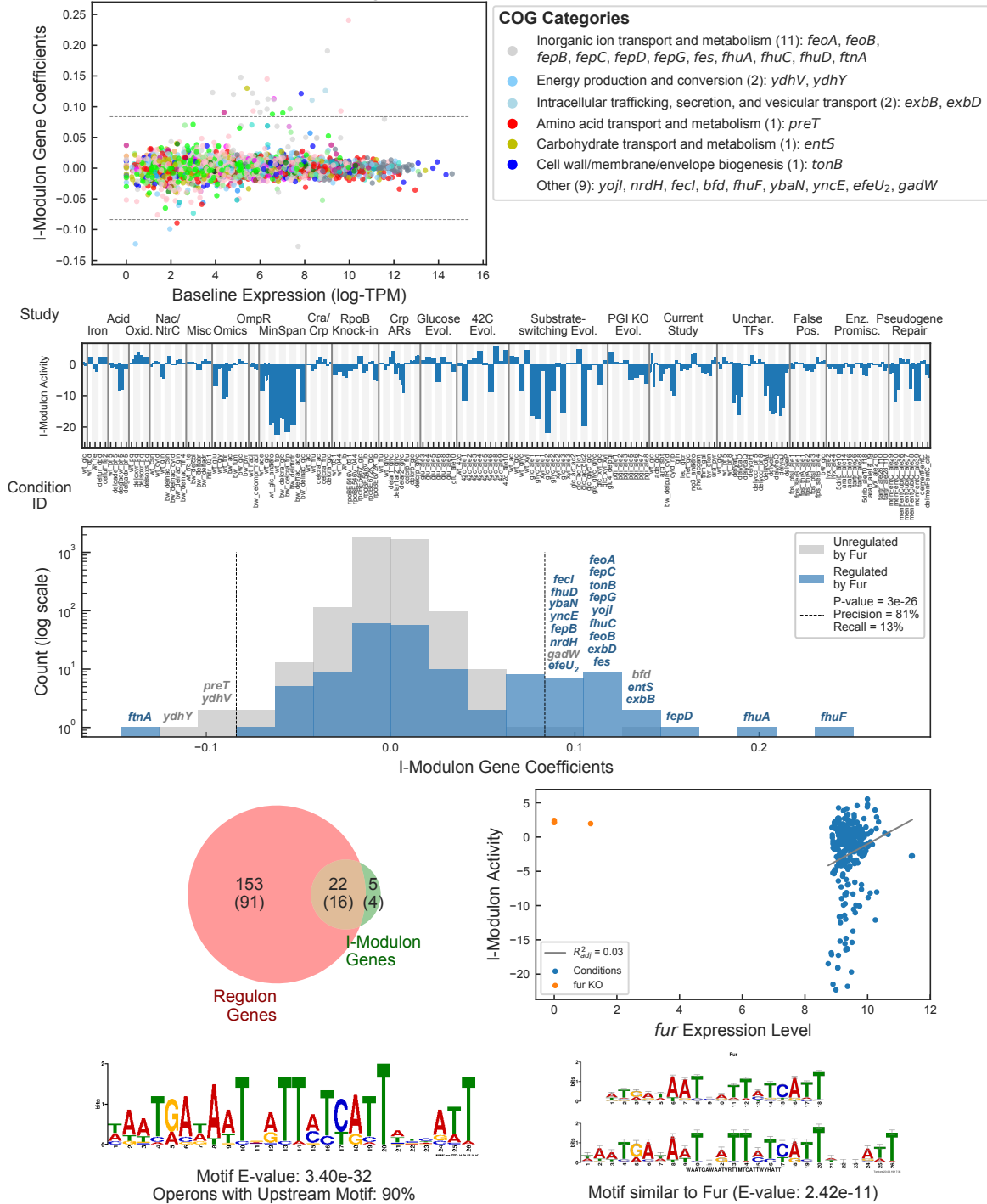
Biological Function: Iron homeostasis



Fur – 2 I-Modulon

Regulated by: Fur

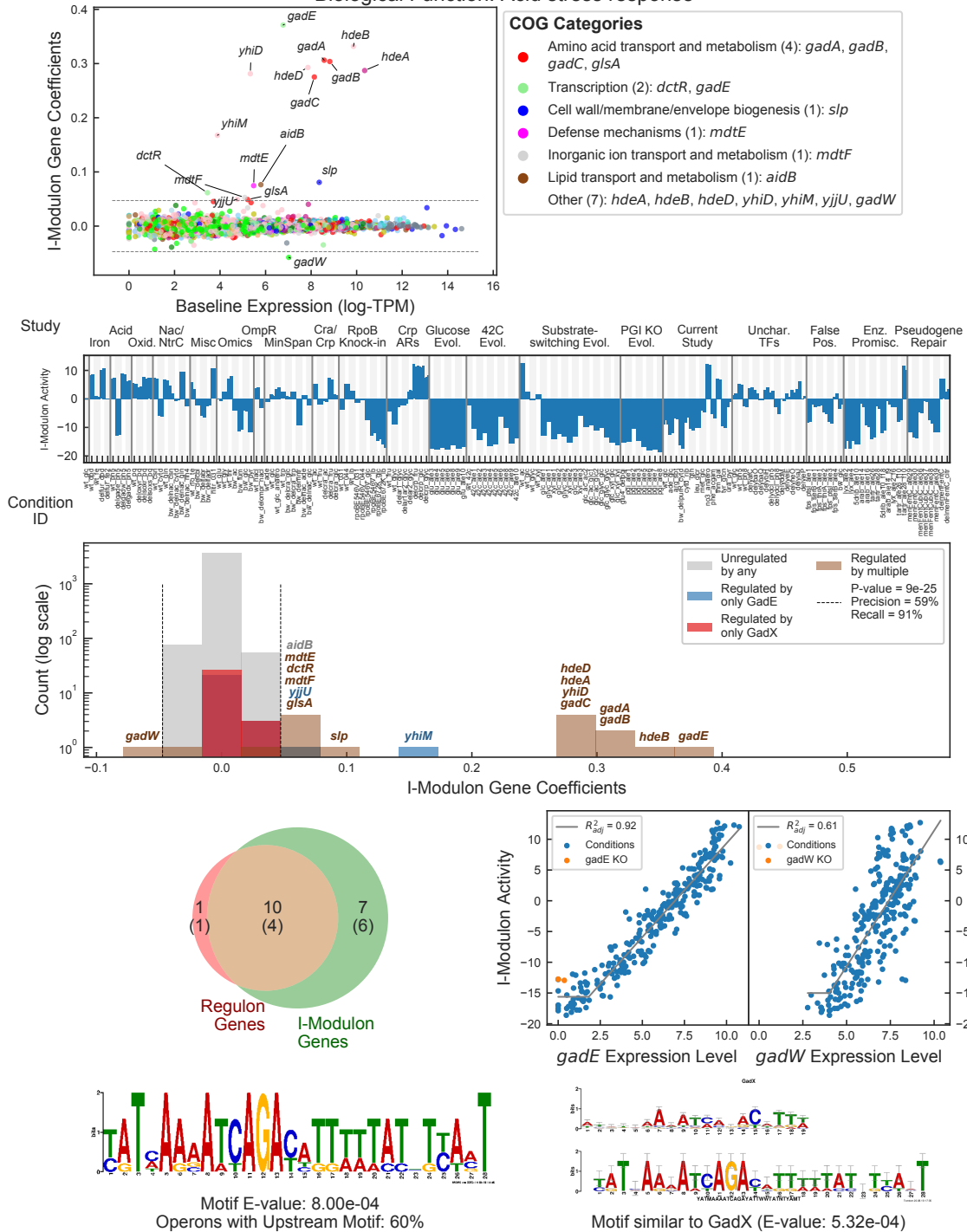
Biological Function: Iron homeostasis



GadEWX I-Modulon

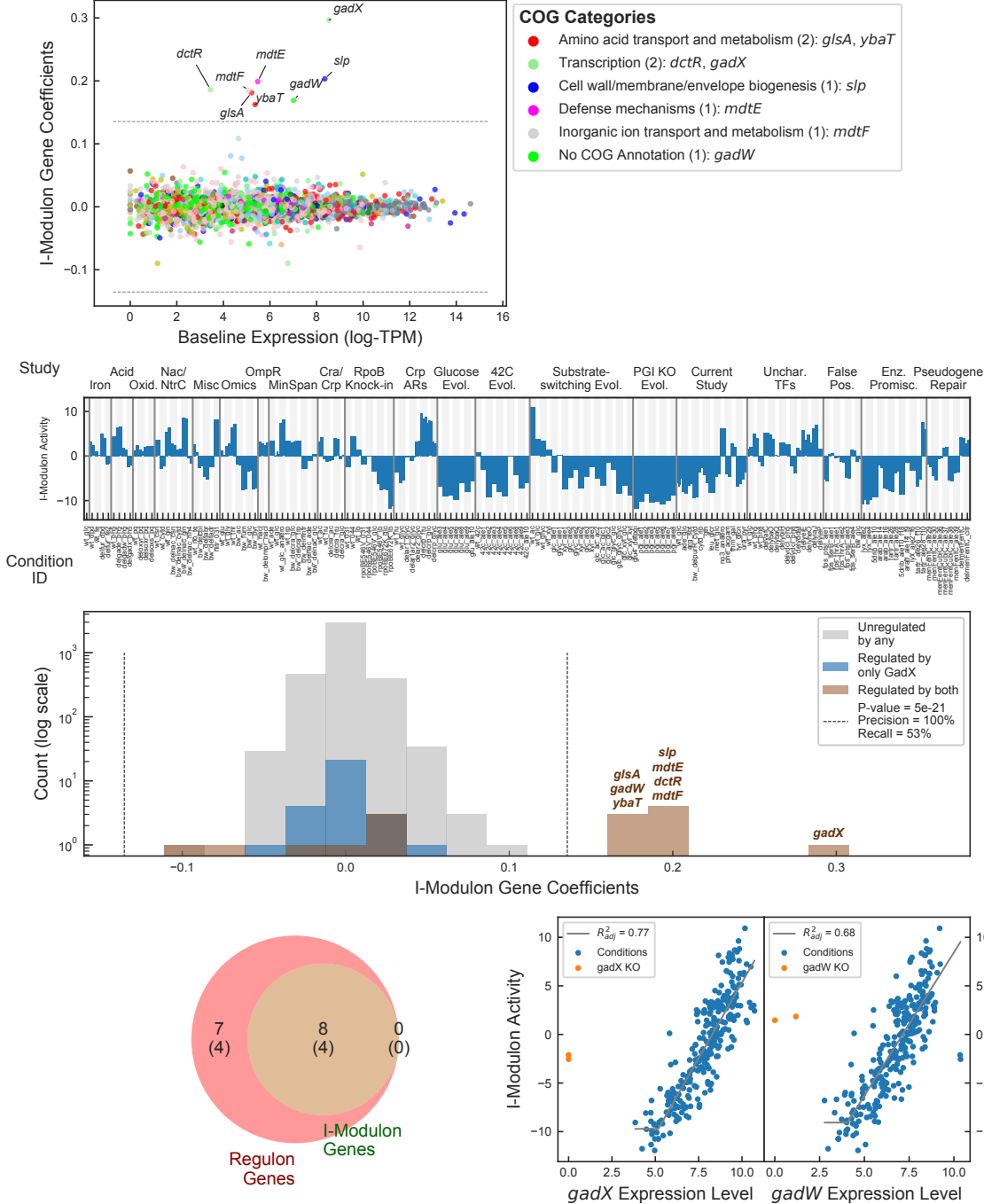
Regulated by: GadE and GadW and GadX

Biological Function: Acid stress response



GadWX I-Modulon

Regulated by: GadW and GadX
Biological Function: Acid stress response



Regulated by: GcvA

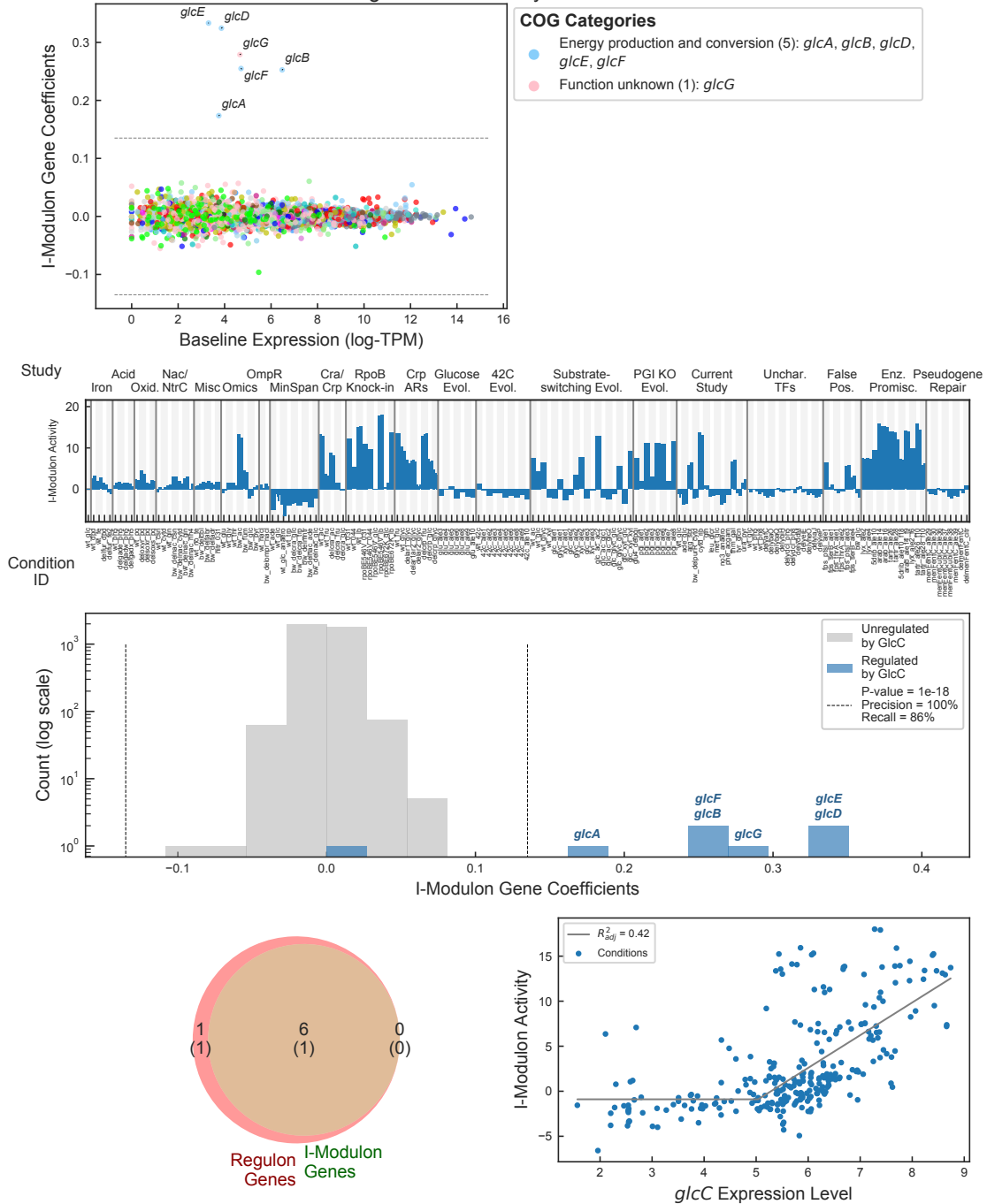
Biological Function: Glycine cleavage system



GlcC I-Modulon

Regulated by: GlcC

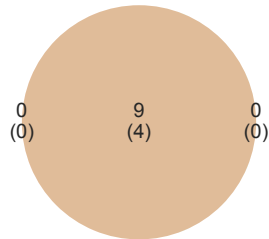
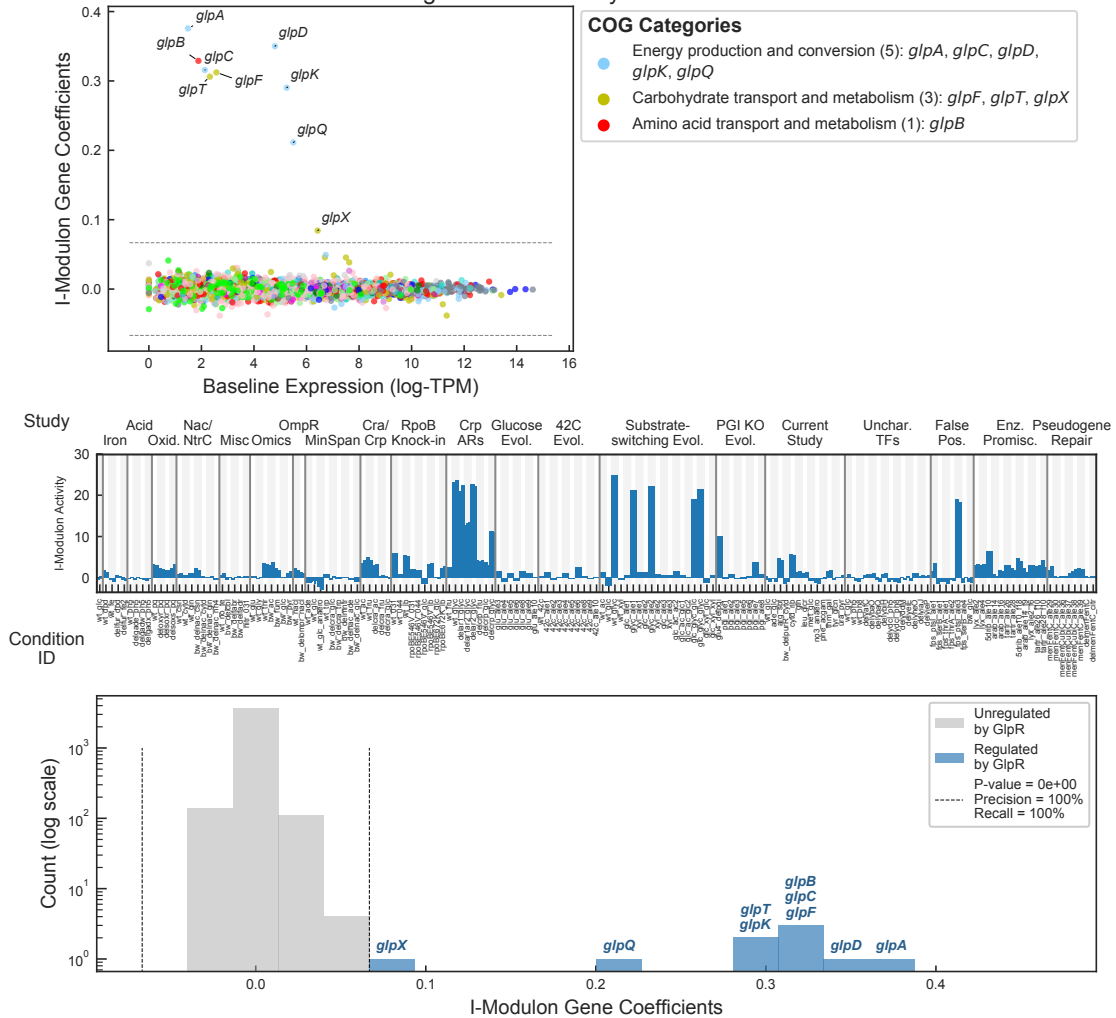
Biological Function: Glycolate catabolism



GlpR I-Modulon

Regulated by: GlpR

Biological Function: Glycerol catabolism

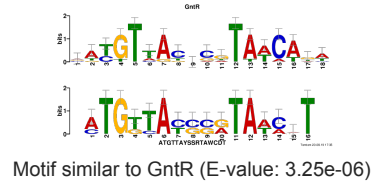
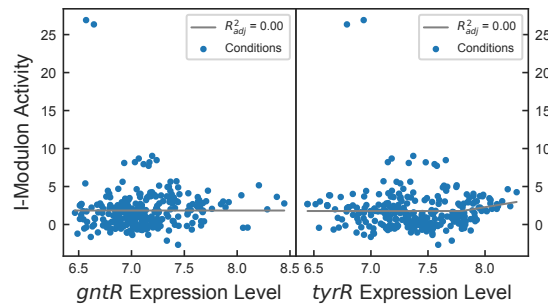
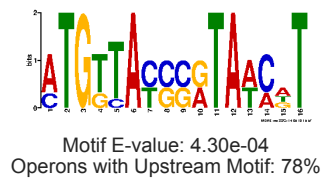
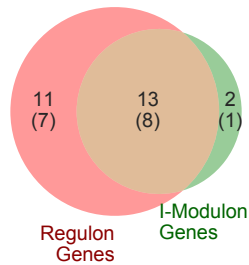
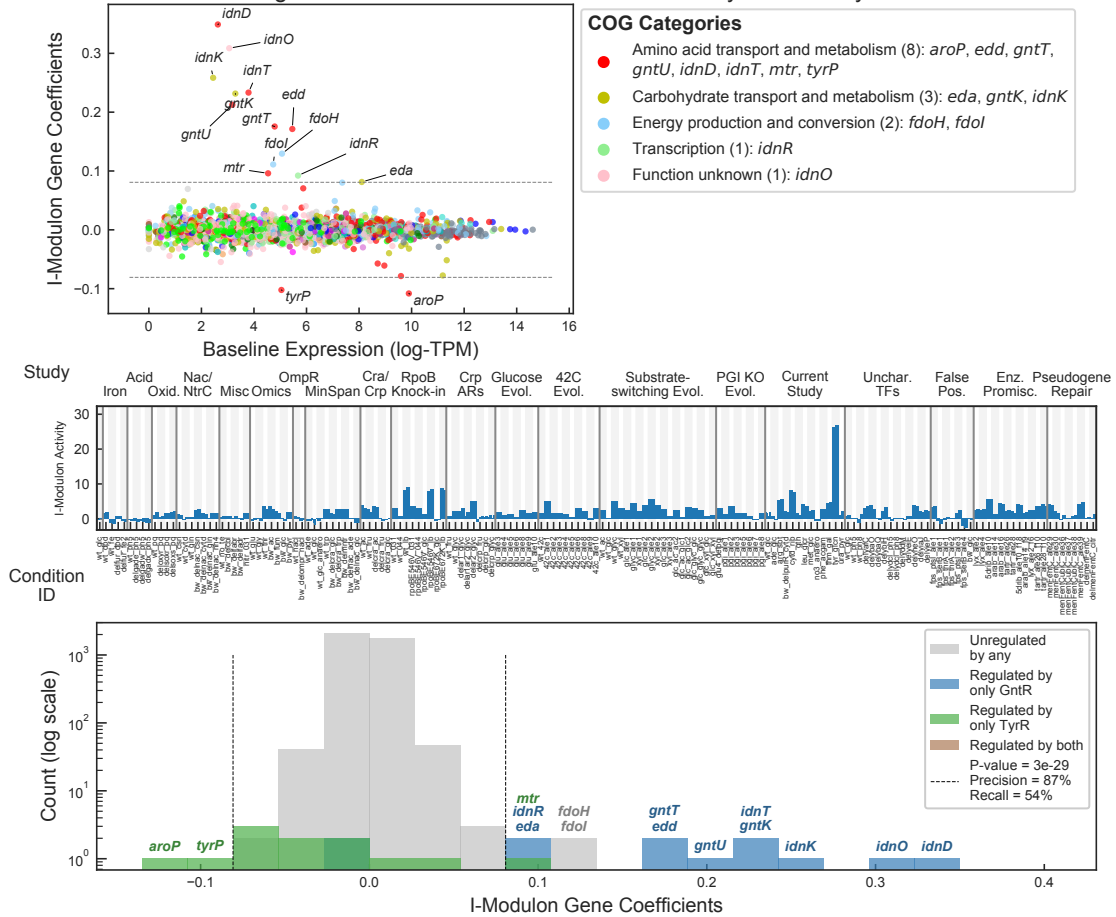


RegulonI-Modulon
GenesGenes

GntR/TyrR I-Modulon

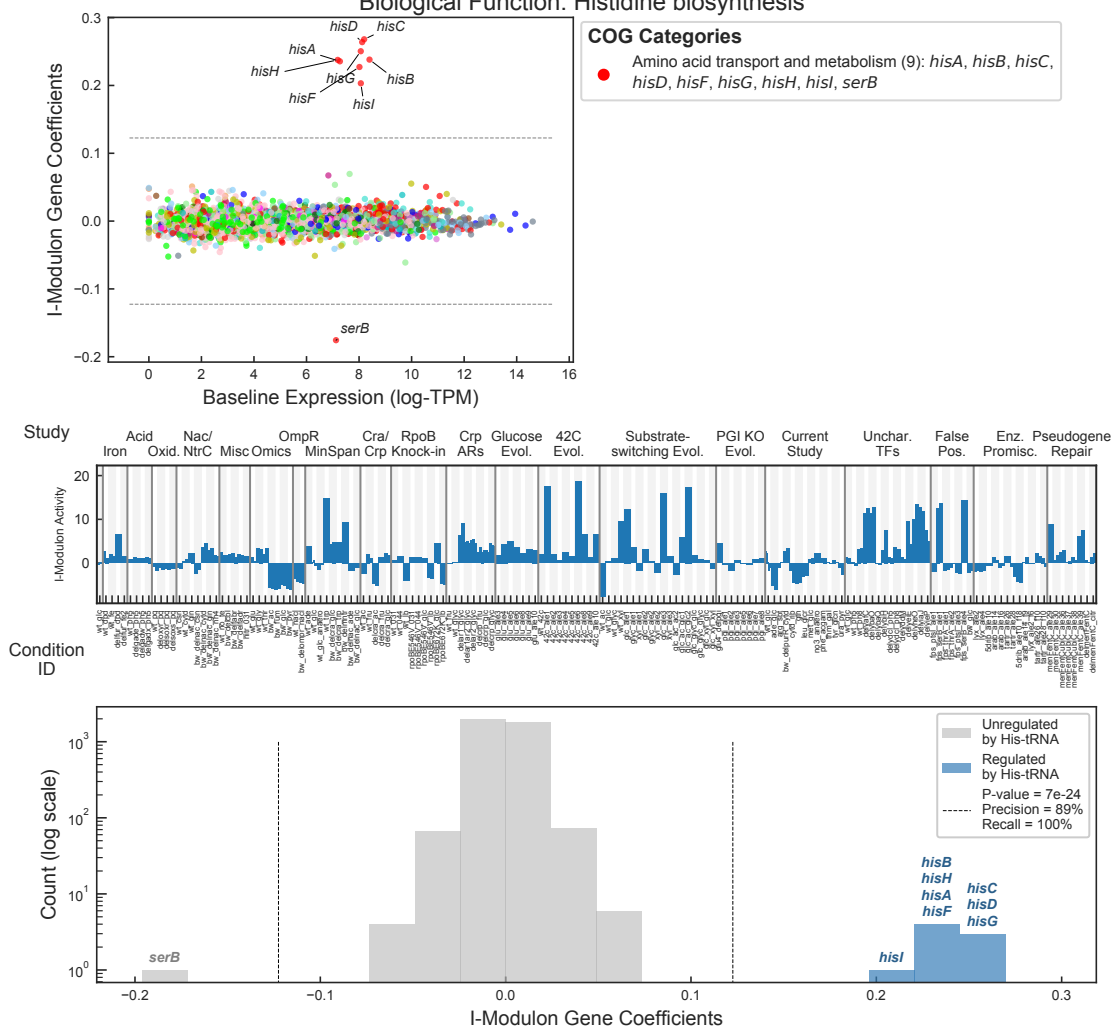
Regulated by: GntR or TyrR

Biological Function: Gluconate catabolism and tyrosine biosynthesis



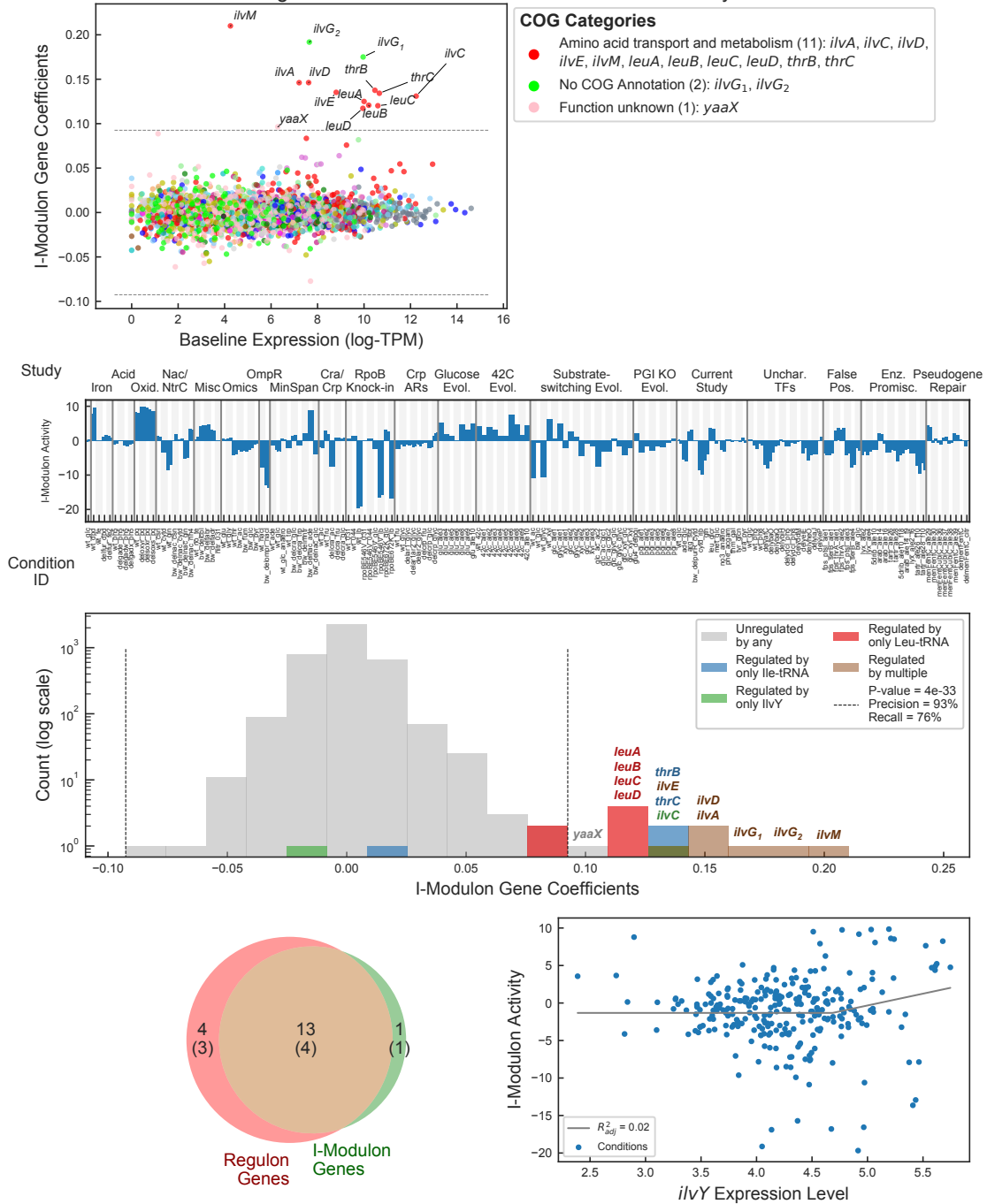
His – tRNA I-Modulon

Regulated by: His-tRNA attenuation
Biological Function: Histidine biosynthesis



Leu/Ile I-Modulon

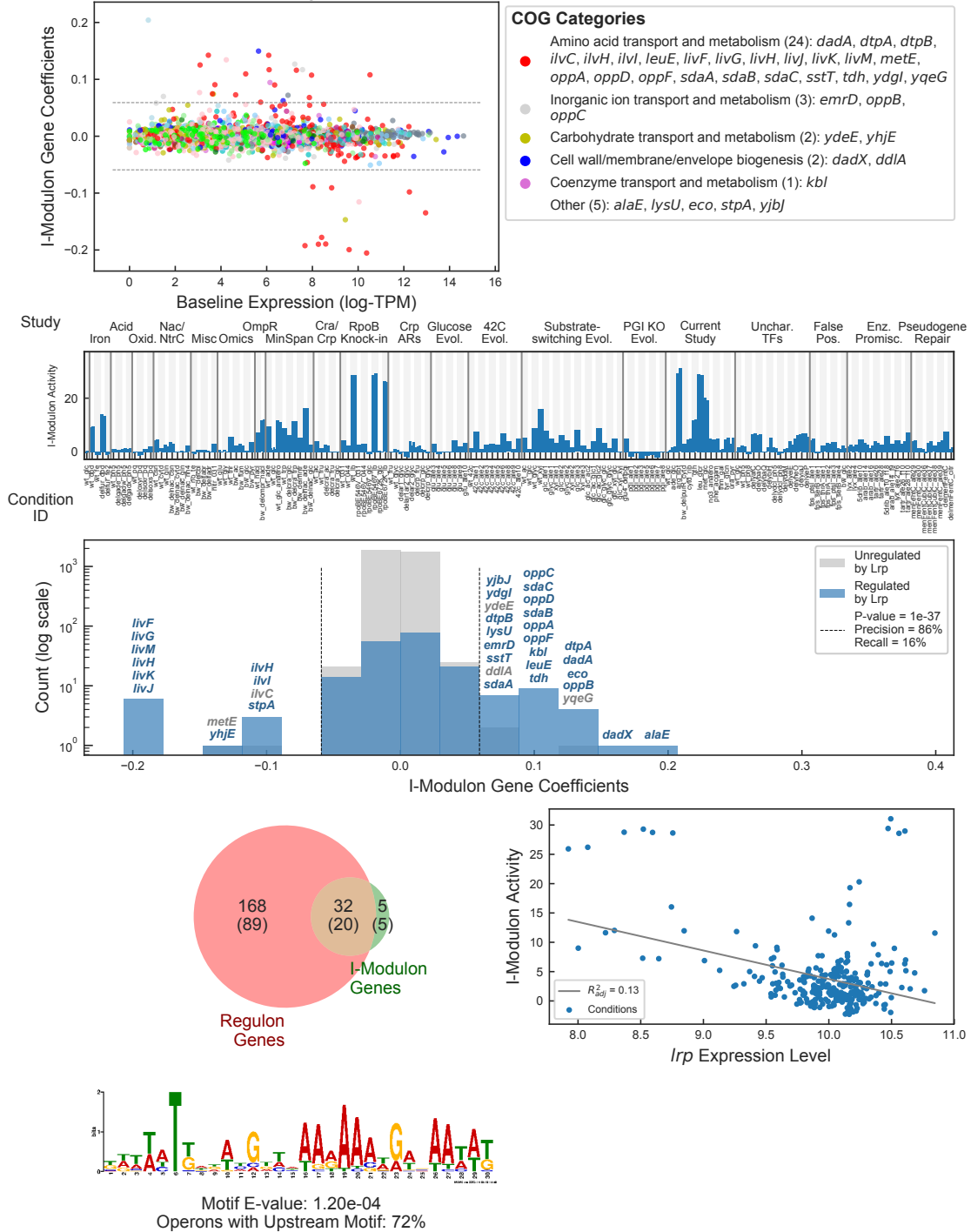
Regulated by: *ilvY* or *leu*-tRNA attenuation or *ile*-tRNA attenuation
Biological Function: Branched-chain amino acid biosynthesis



Lrp I-Modulon

Regulated by: Lrp

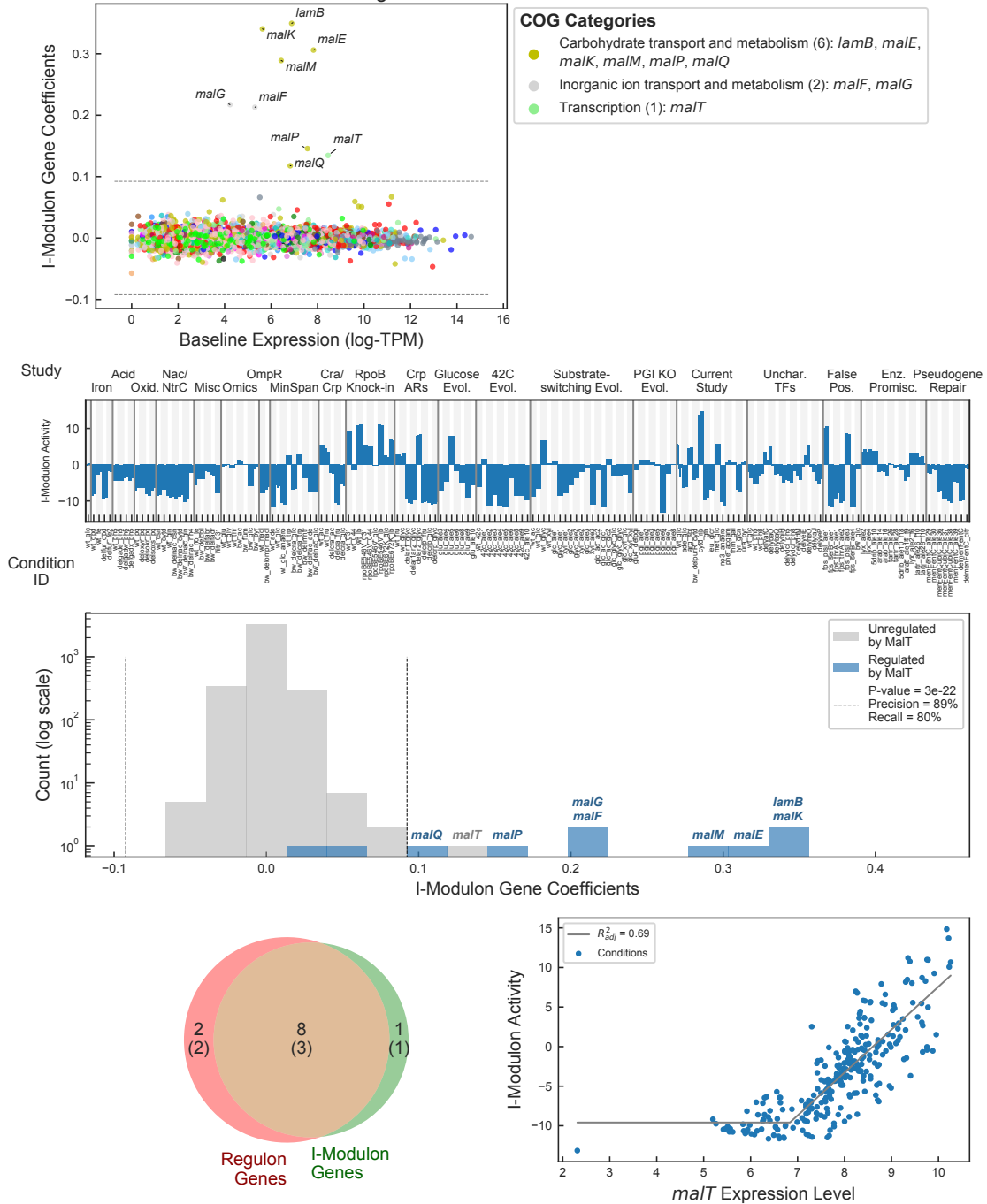
Biological Function: Amino acid and peptide transport



MalT I-Modulon

Regulated by: MalT

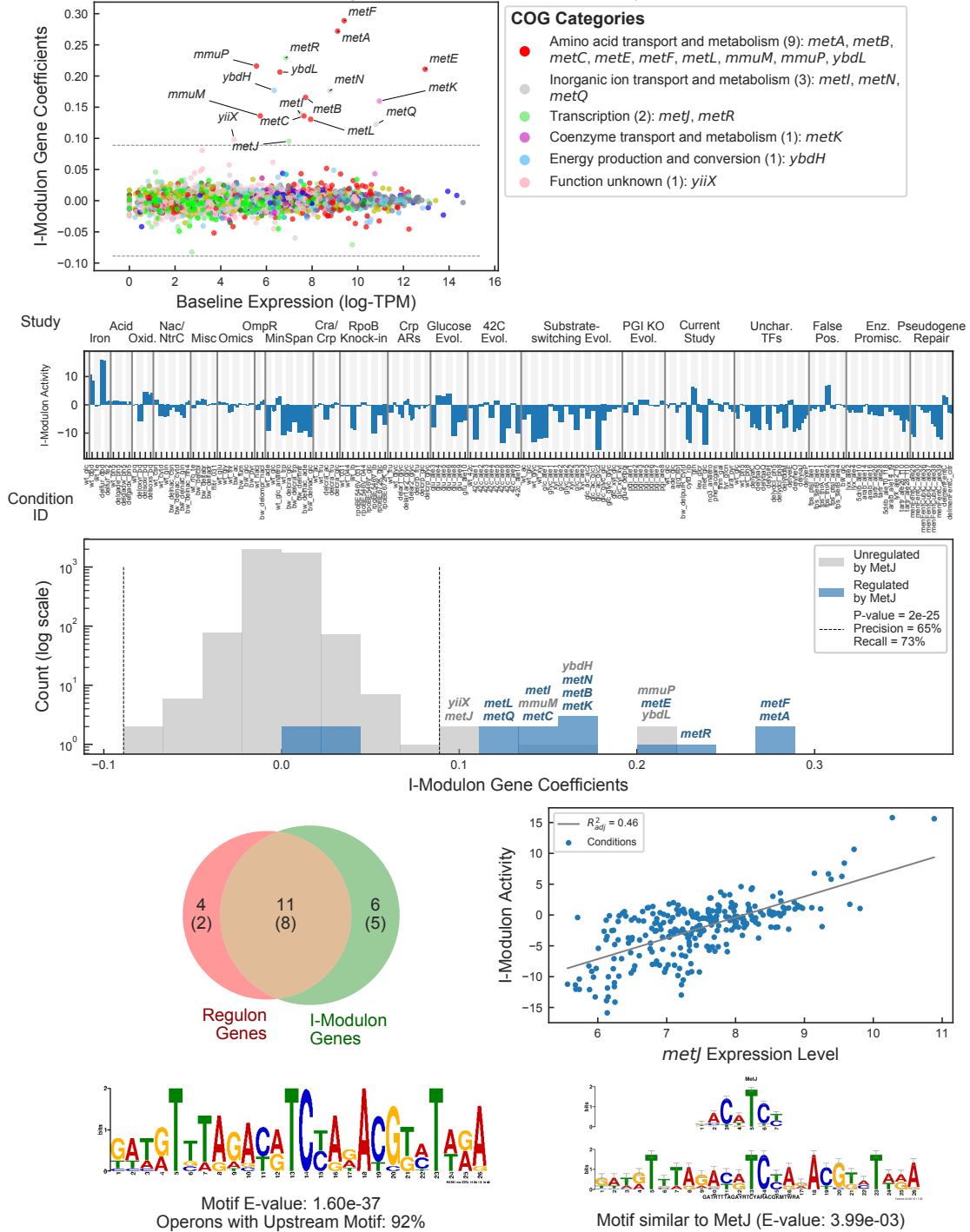
Biological Function: Maltose catabolism



MetJ I-Modulon

Regulated by: MetJ

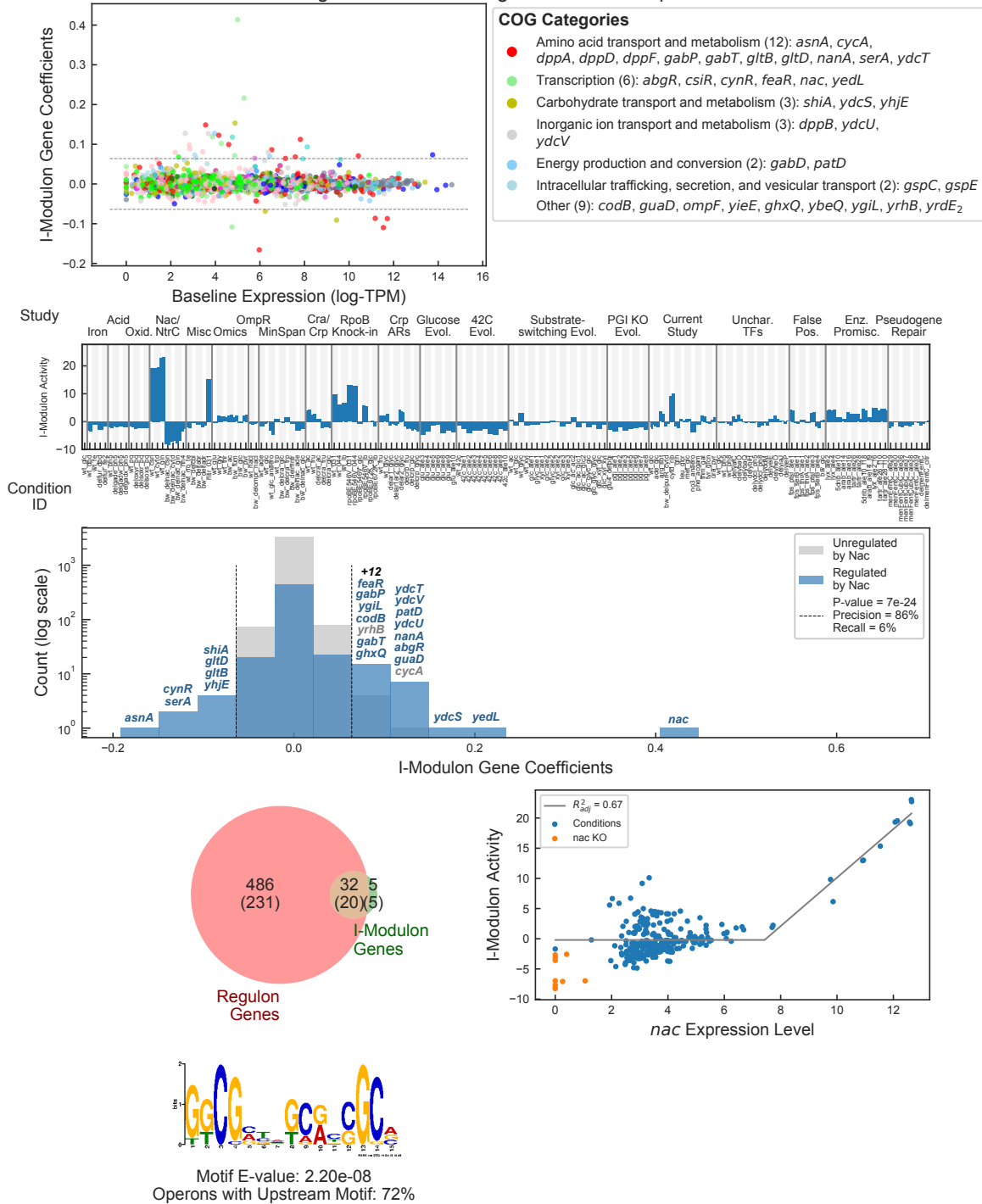
Biological Function: Methionine biosynthesis



Nac I-Modulon

Regulated by: Nac

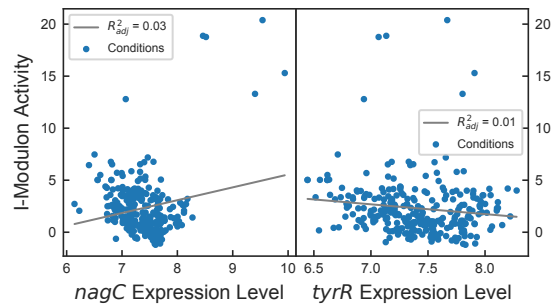
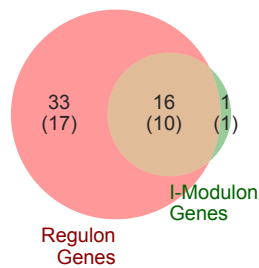
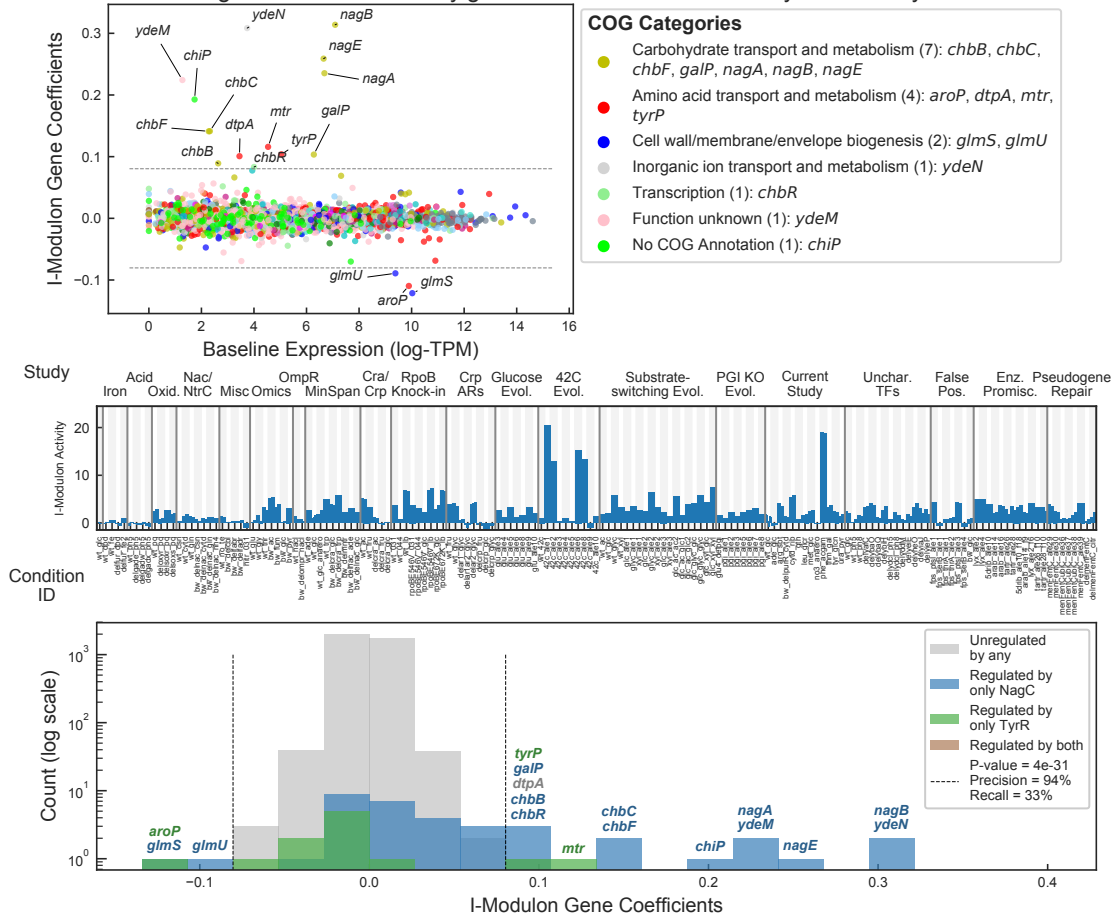
Biological Function: Nitrogen starvation response



NagC/TyrR I-Modulon

Regulated by: NagC or TyrR

Biological Function: N-acetylglucosamine catabolism and tyrosine biosynthesis

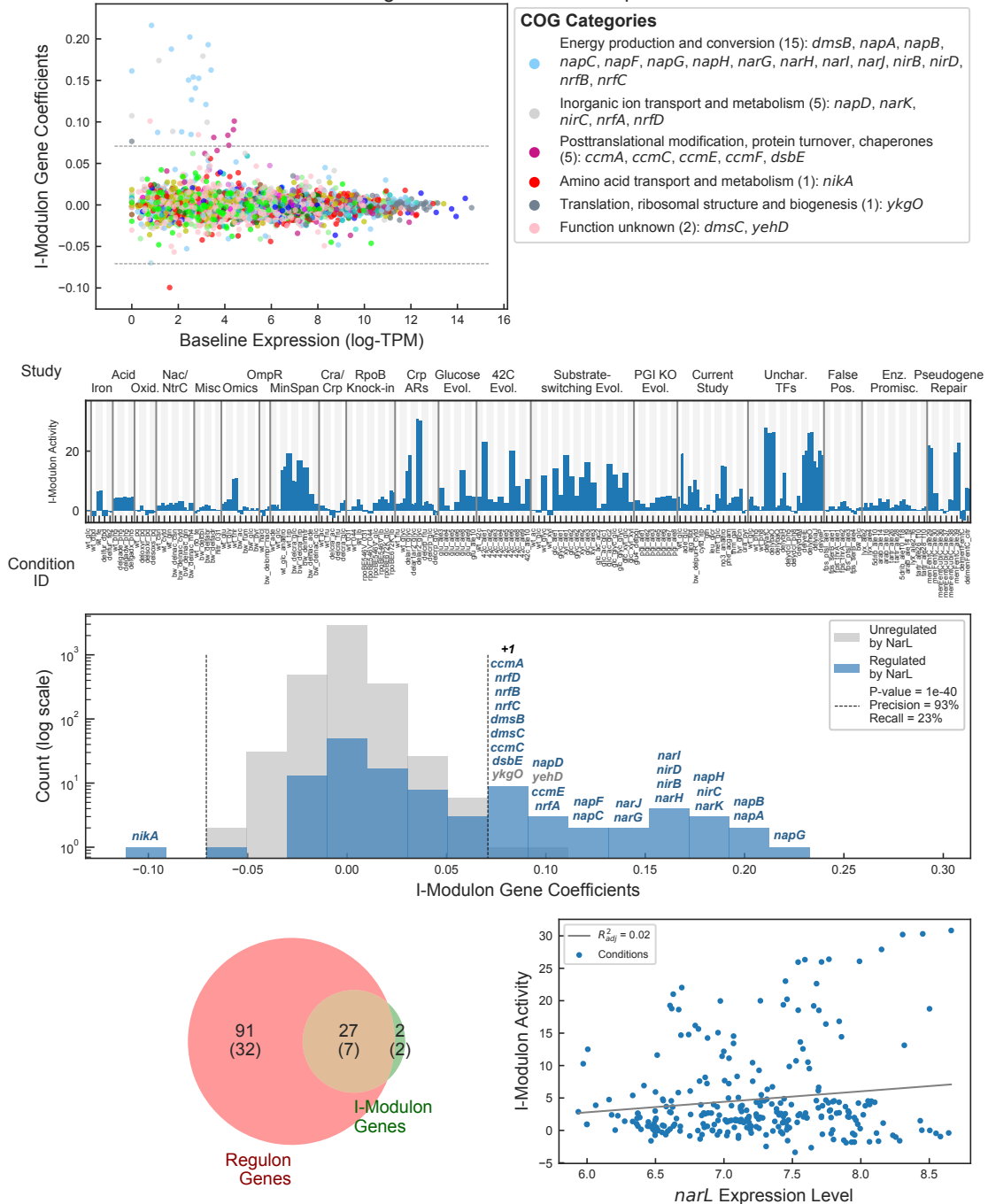


Motif E-value: 1.50e-04
Operons with Upstream Motif: 27%

NarL I-Modulon

Regulated by: NarL

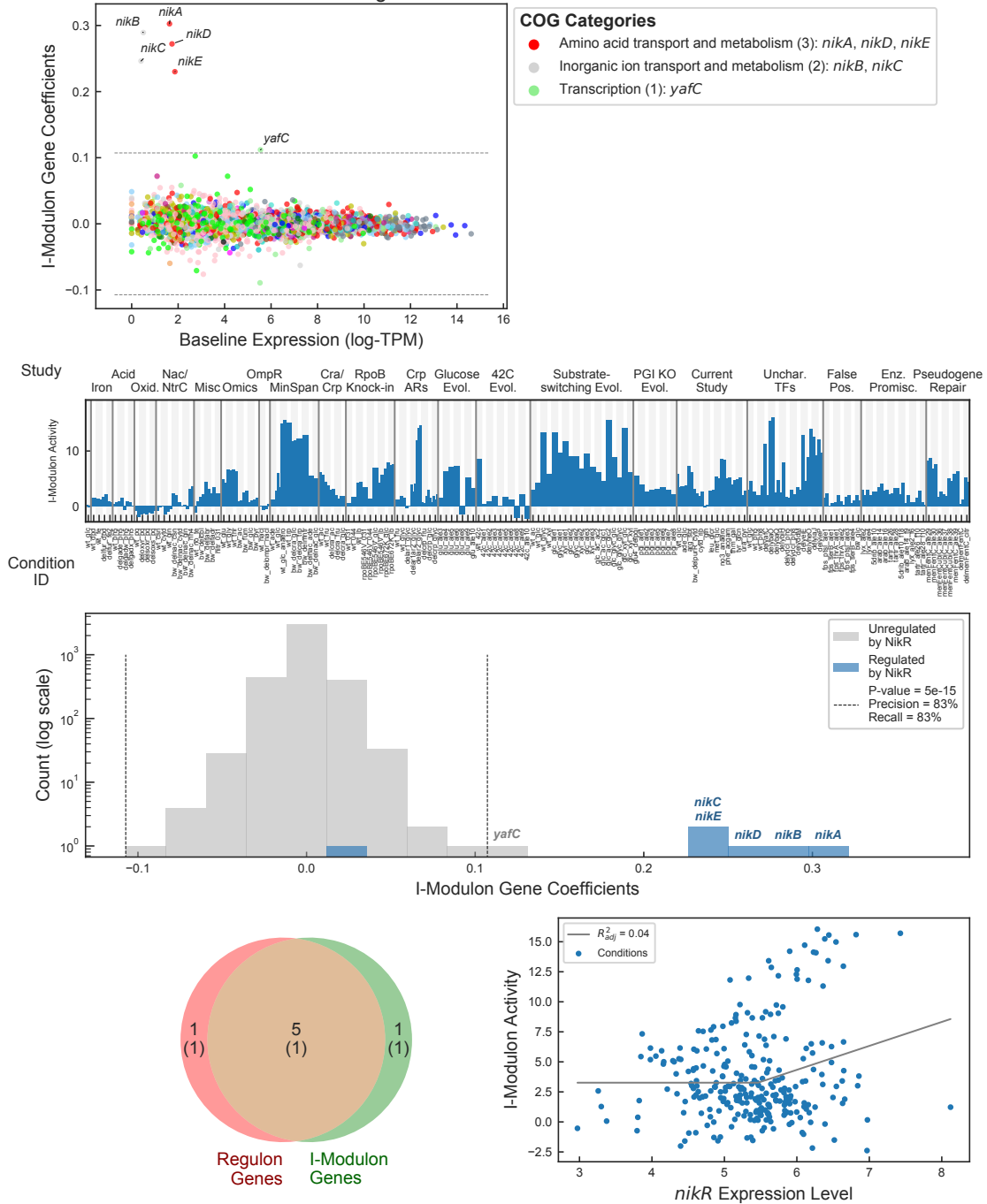
Biological Function: Nitrate respiration



NikR I-Modulon

Regulated by: NikR

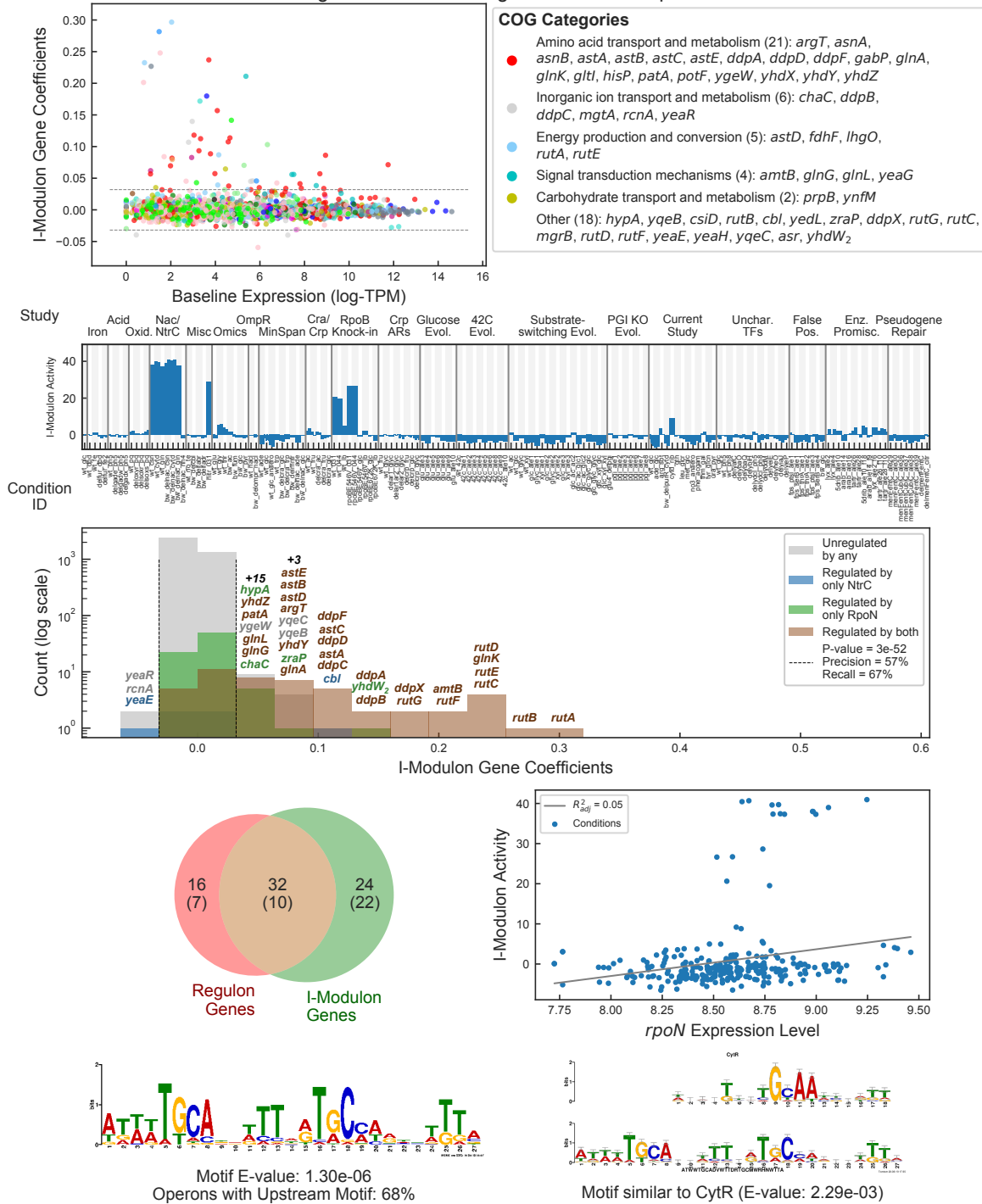
Biological Function: Nickel homeostasis



NtrC + RpoN I-Modulon

Regulated by: NtrC and RpoN

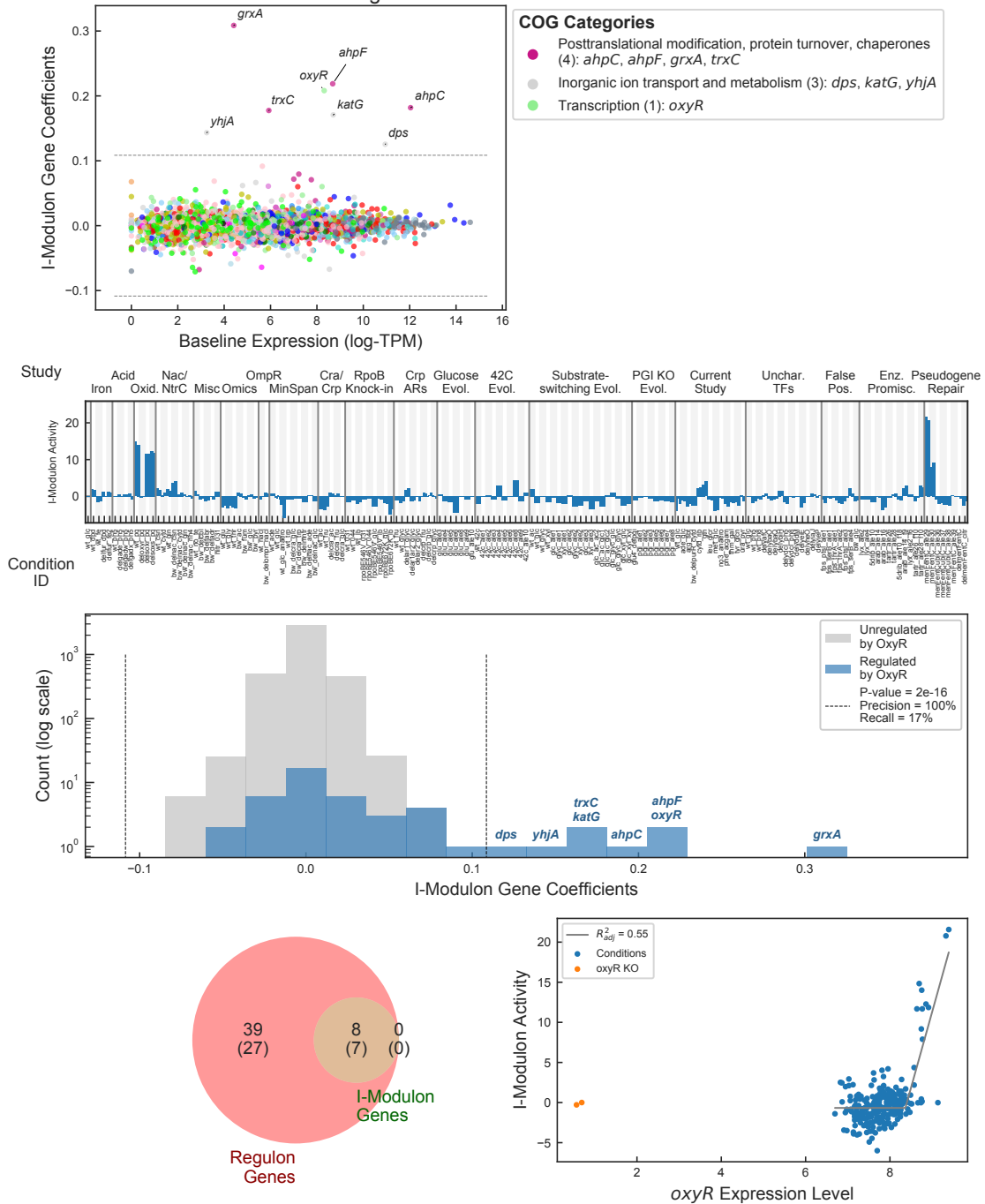
Biological Function: Nitrogen starvation response



OxyR I-Modulon

Regulated by: OxyR

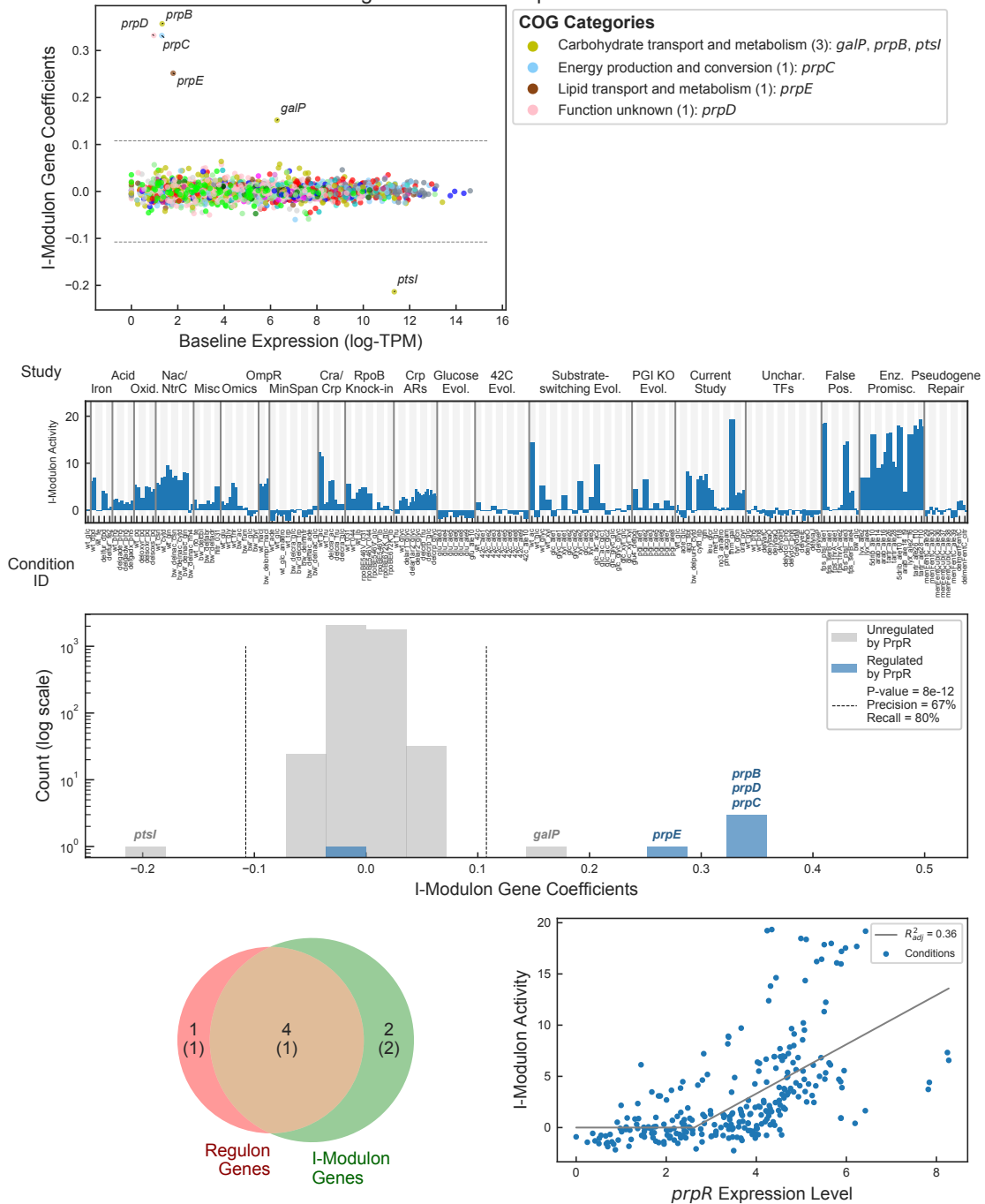
Biological Function: Peroxide reductases



PrpR I-Modulon

Regulated by: PrpR

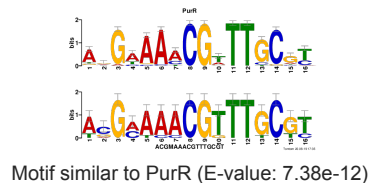
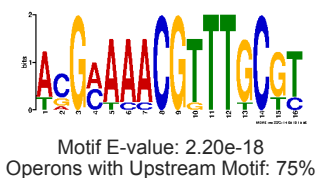
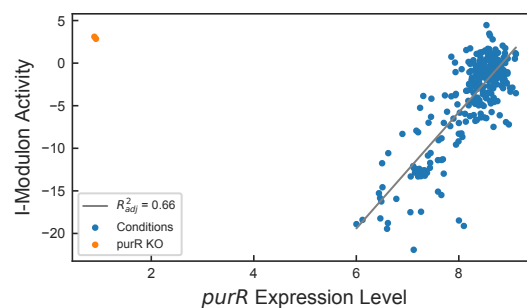
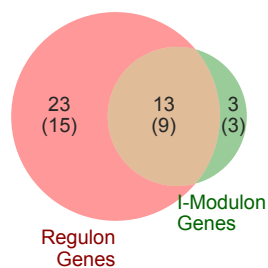
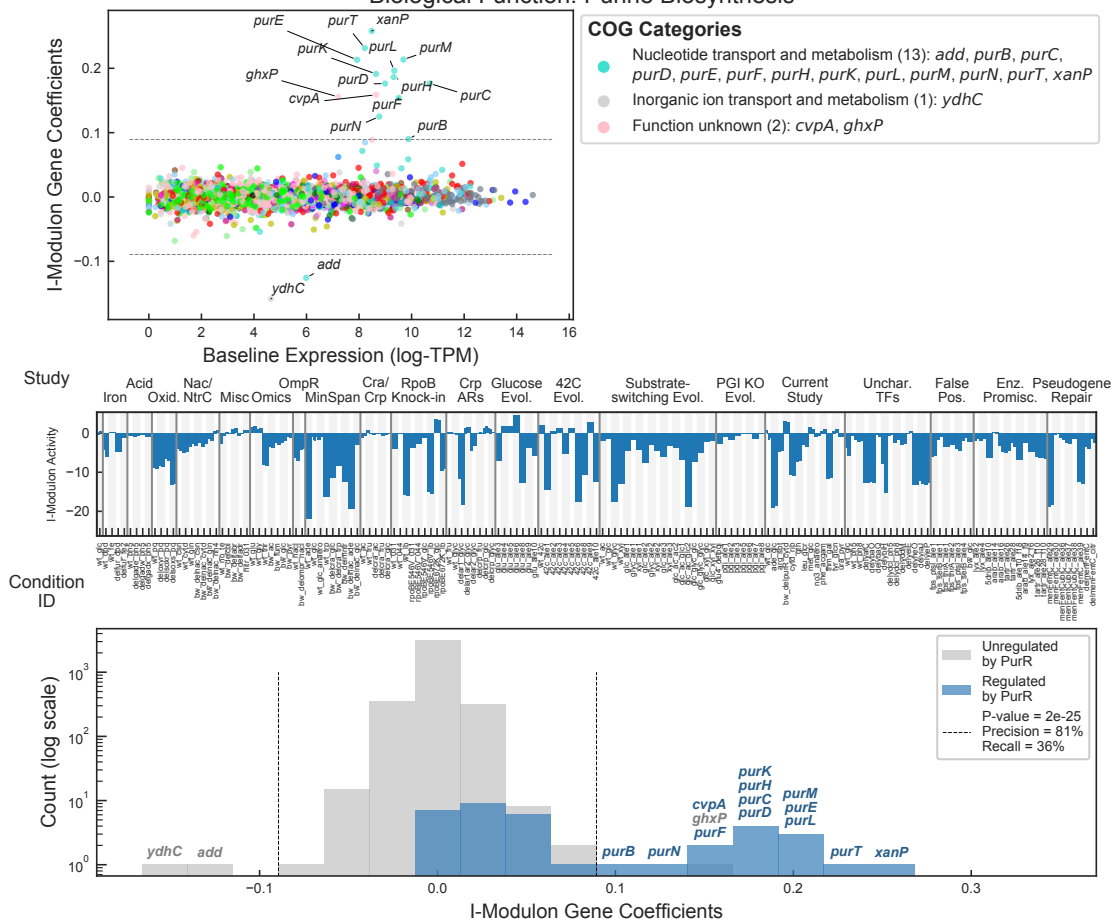
Biological Function: Propionate catabolism



PurR – 1 I-Modulon

Regulated by: PurR

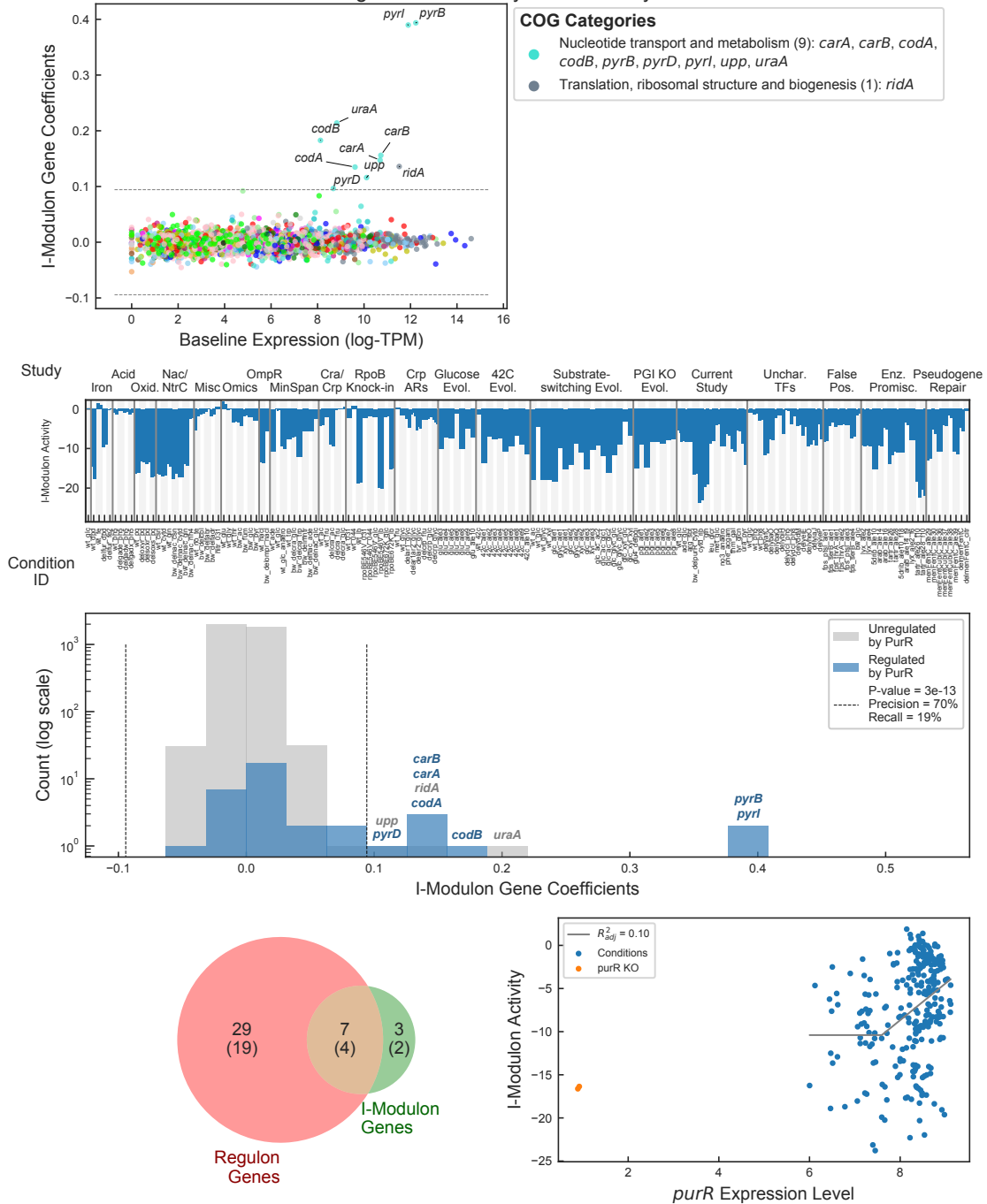
Biological Function: Purine Biosynthesis



PurR – 2 I-Modulon

Regulated by: PurR

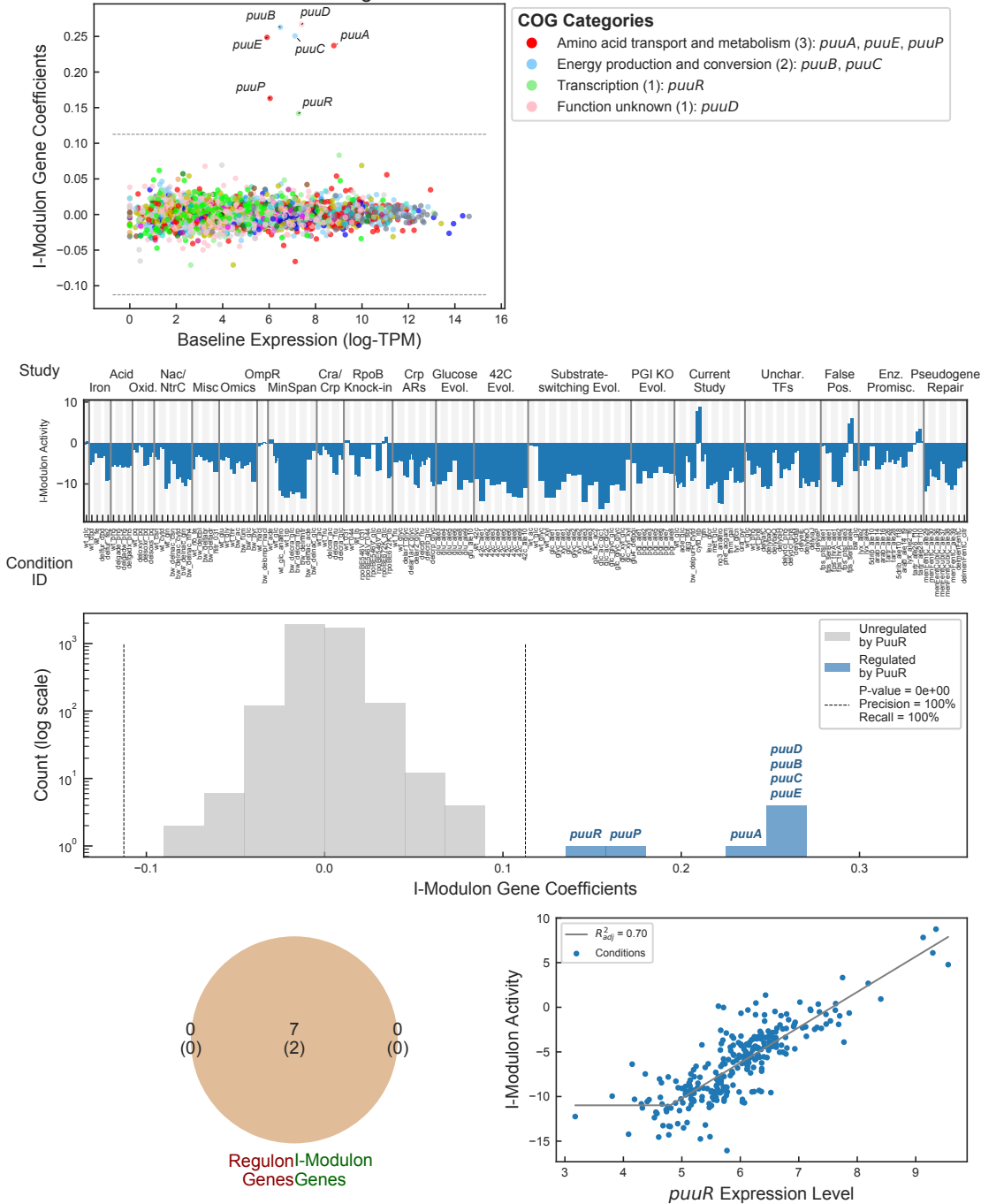
Biological Function: Pyrimidine biosynthesis



PuuR I-Modulon

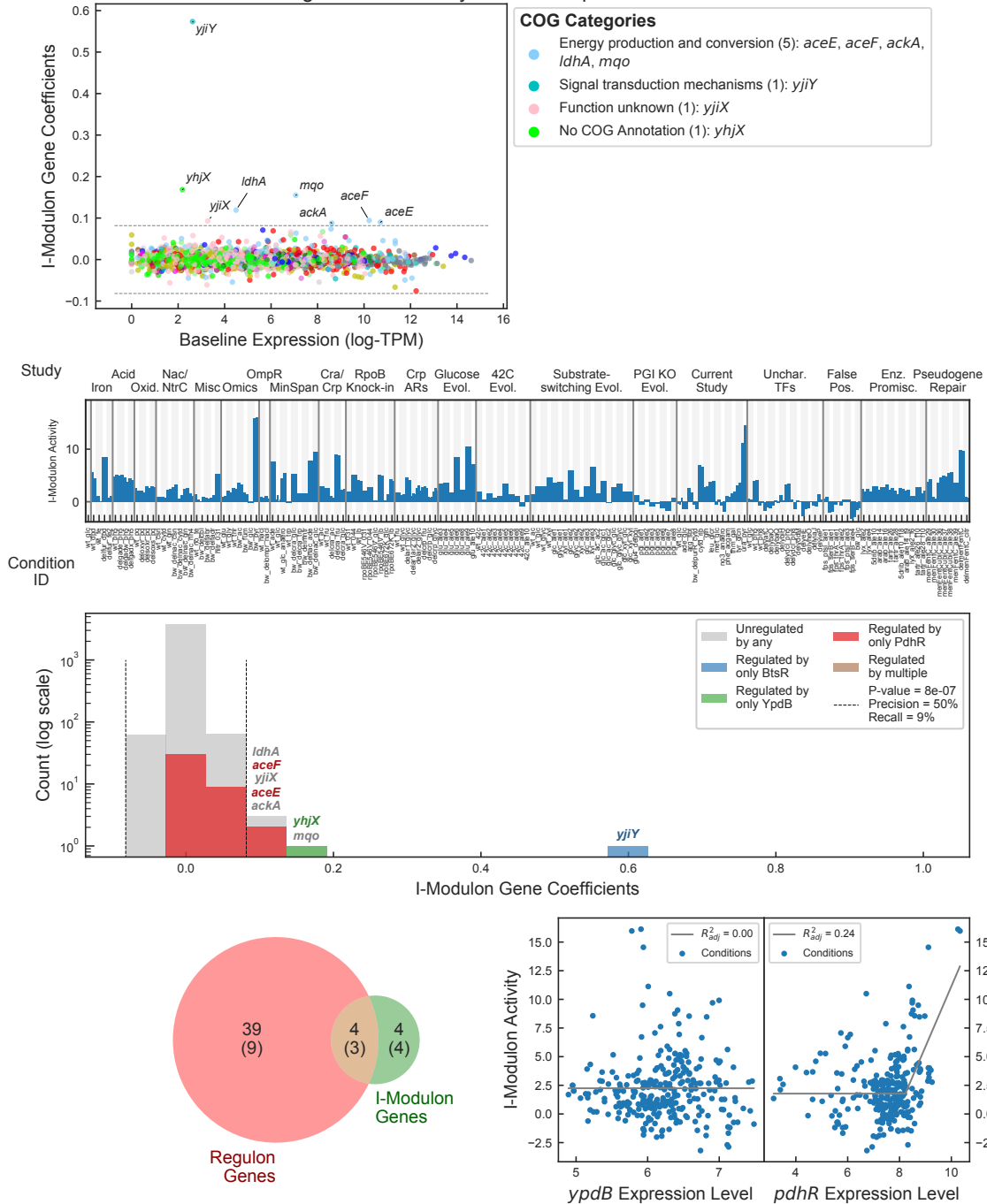
Regulated by: PuuR

Biological Function: Putrescine catabolism



Pyruvate I-Modulon

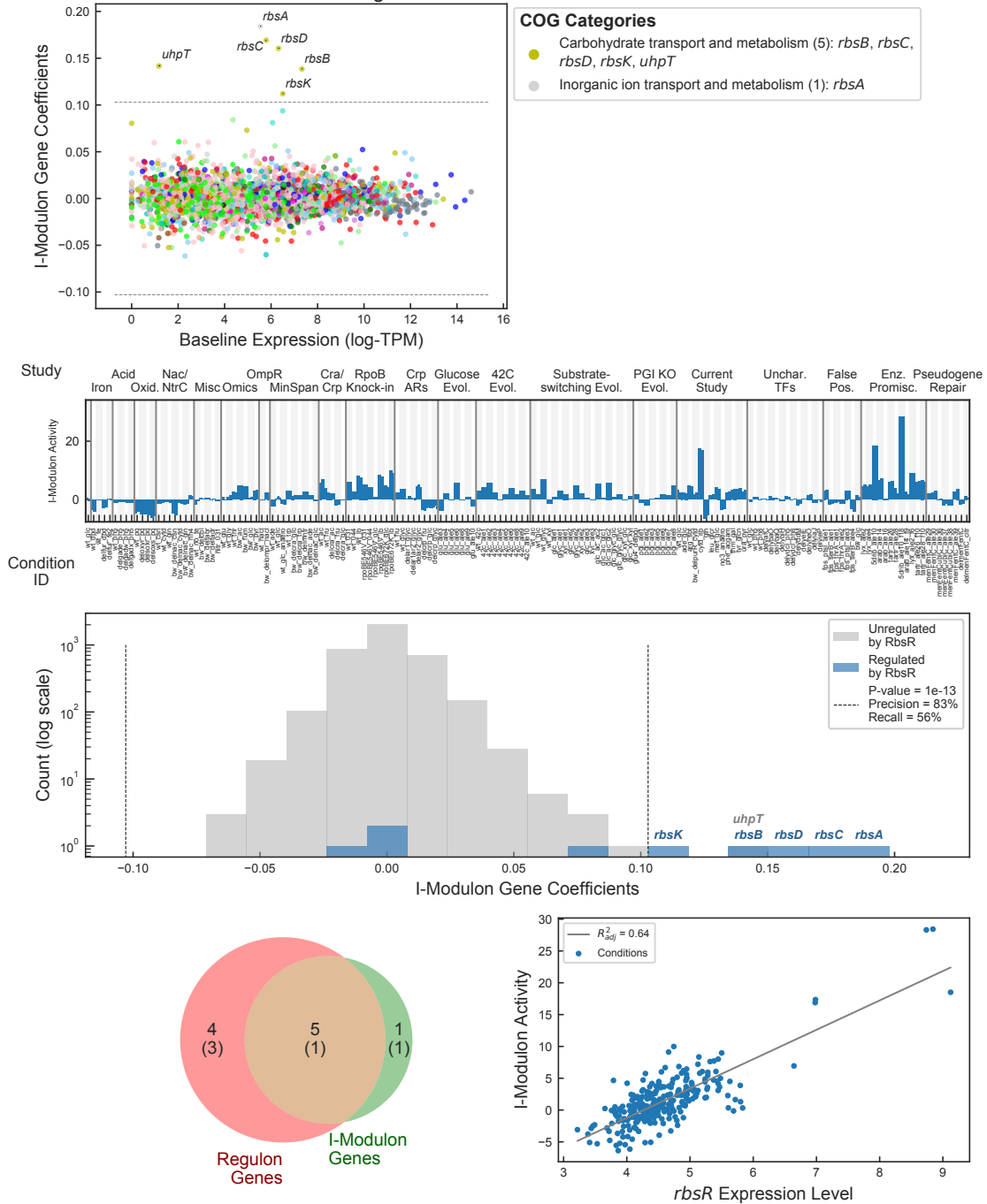
Regulated by: BtsR or YpdB or PdhR
Biological Function: Pyruvate transport and metabolism



RbsR I-Modulon

Regulated by: RbsR

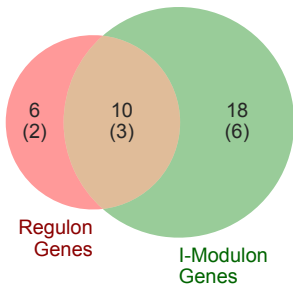
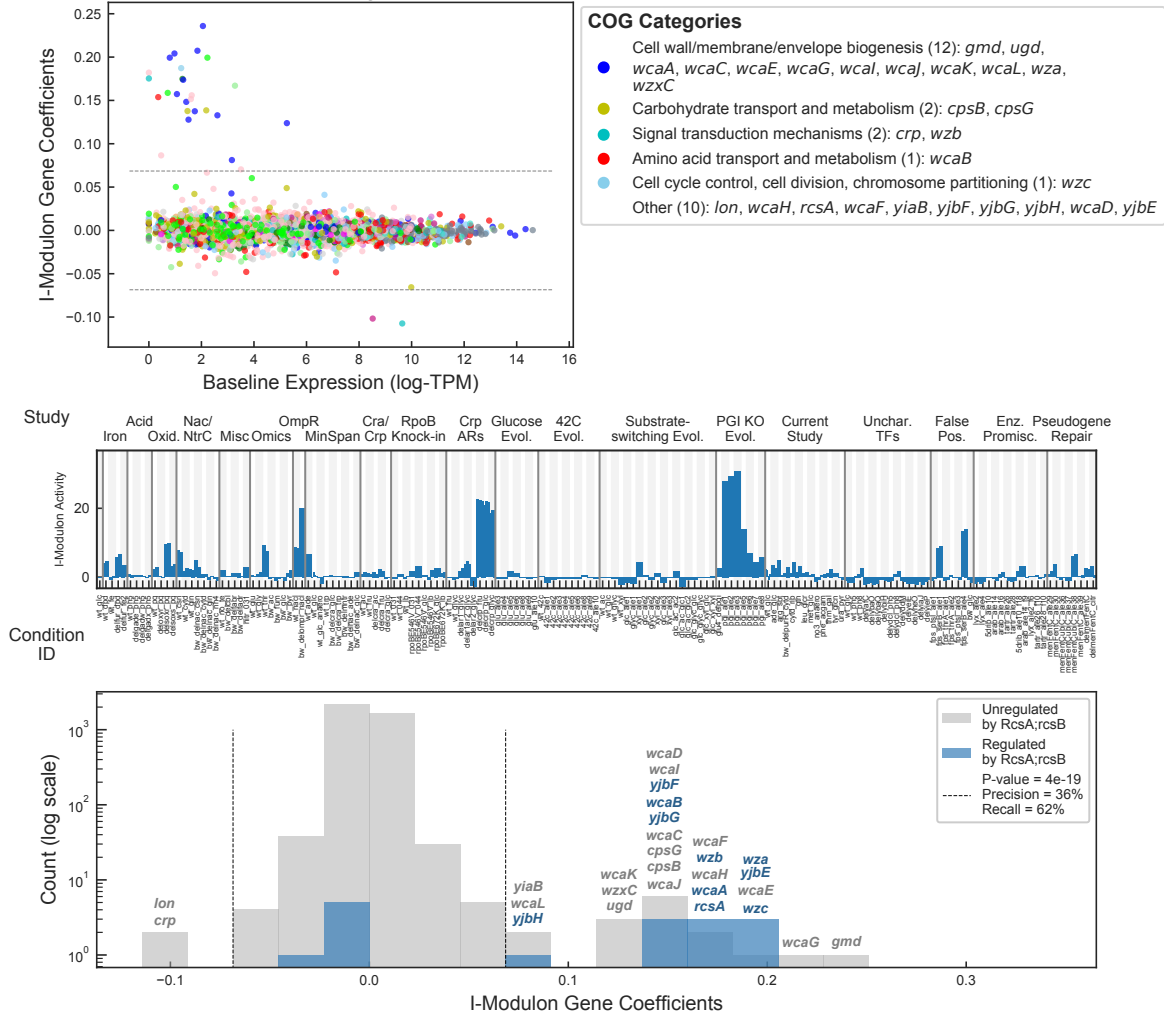
Biological Function: D-ribose catabolism



RcsAB I-Modulon

Regulated by: RcsAB

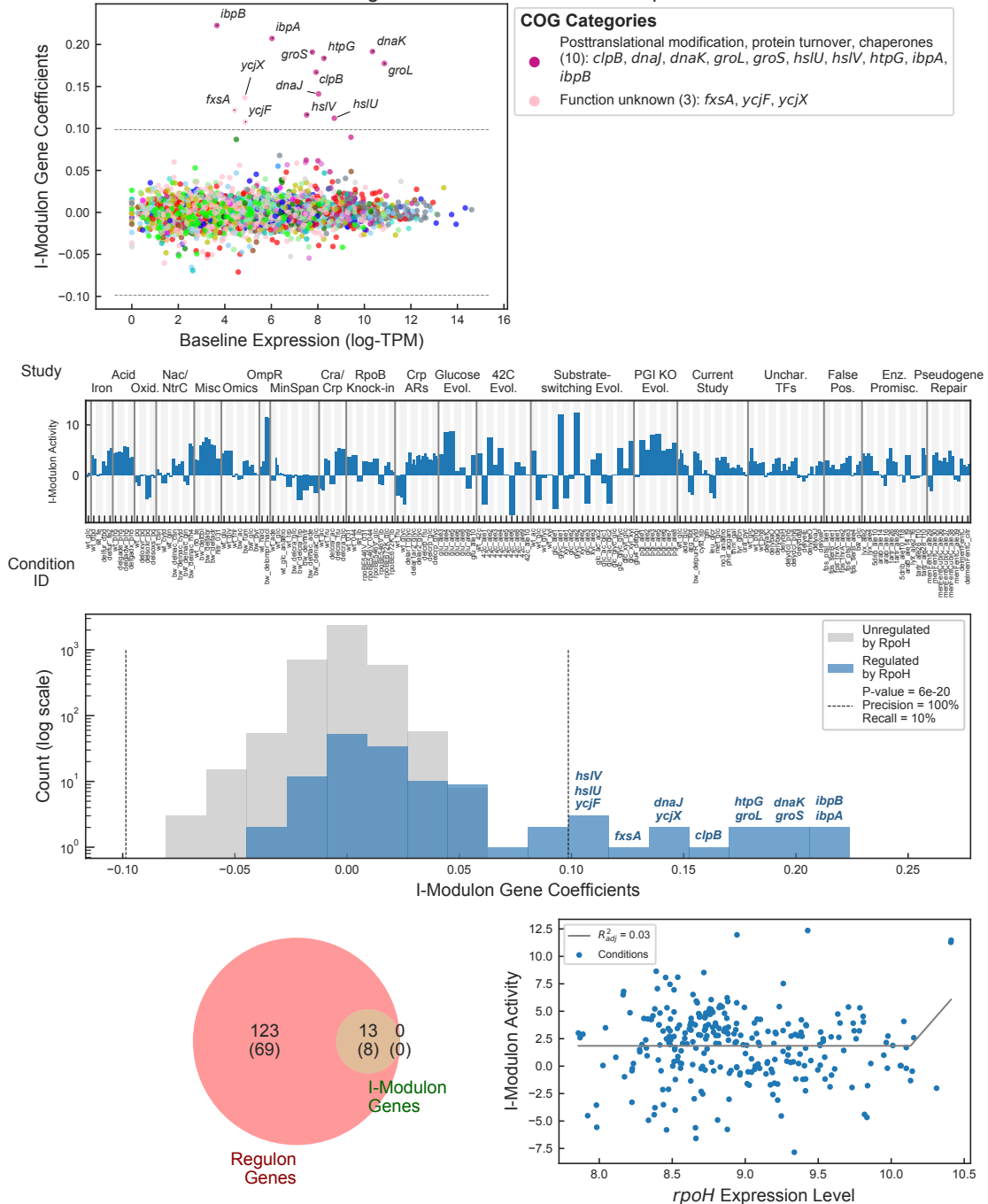
Biological Function: Colanic acid capsule formation



RpoH I-Modulon

Regulated by: RpoH

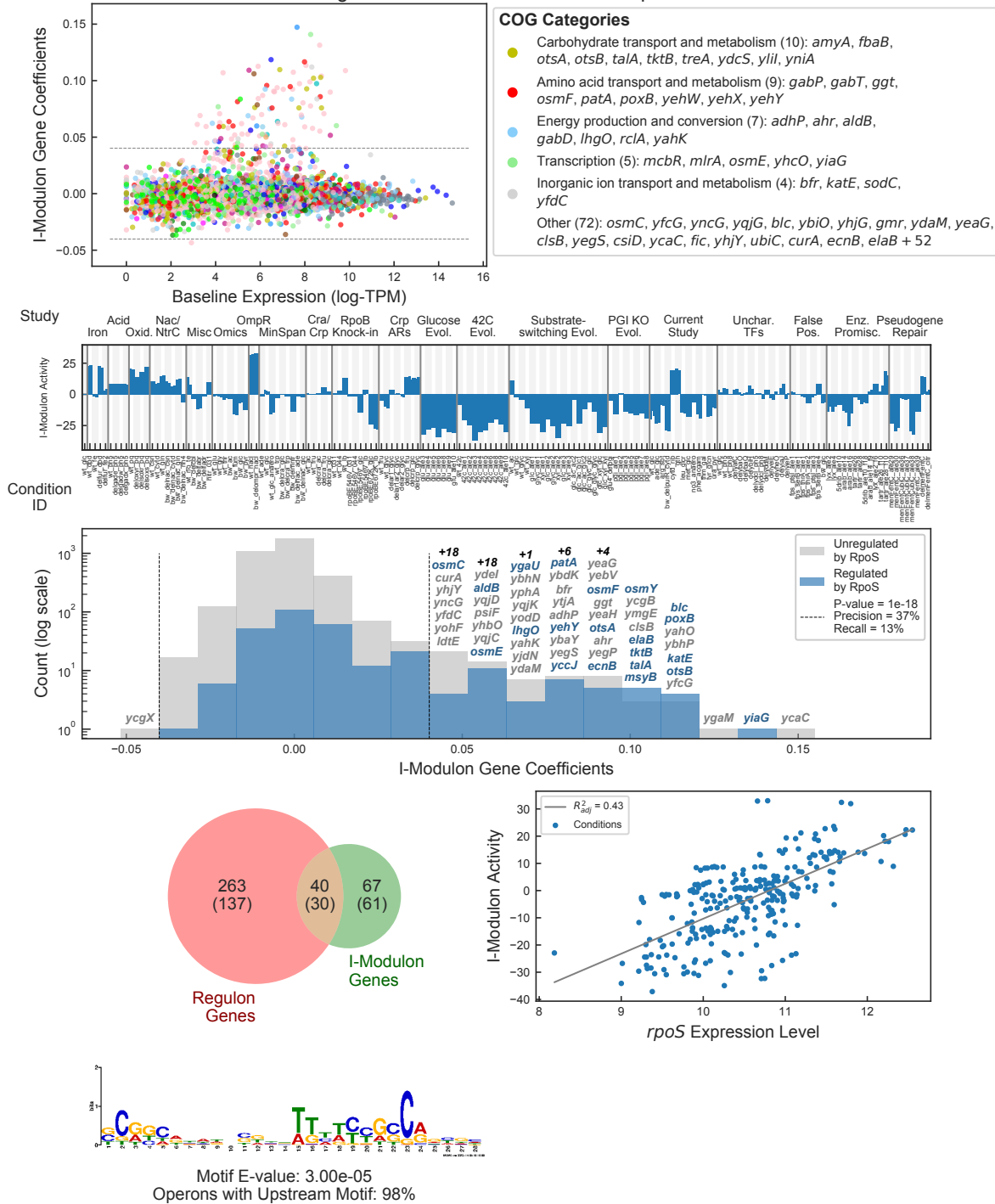
Biological Function: Heat shock response



RpoS I-Modulon

Regulated by: RpoS

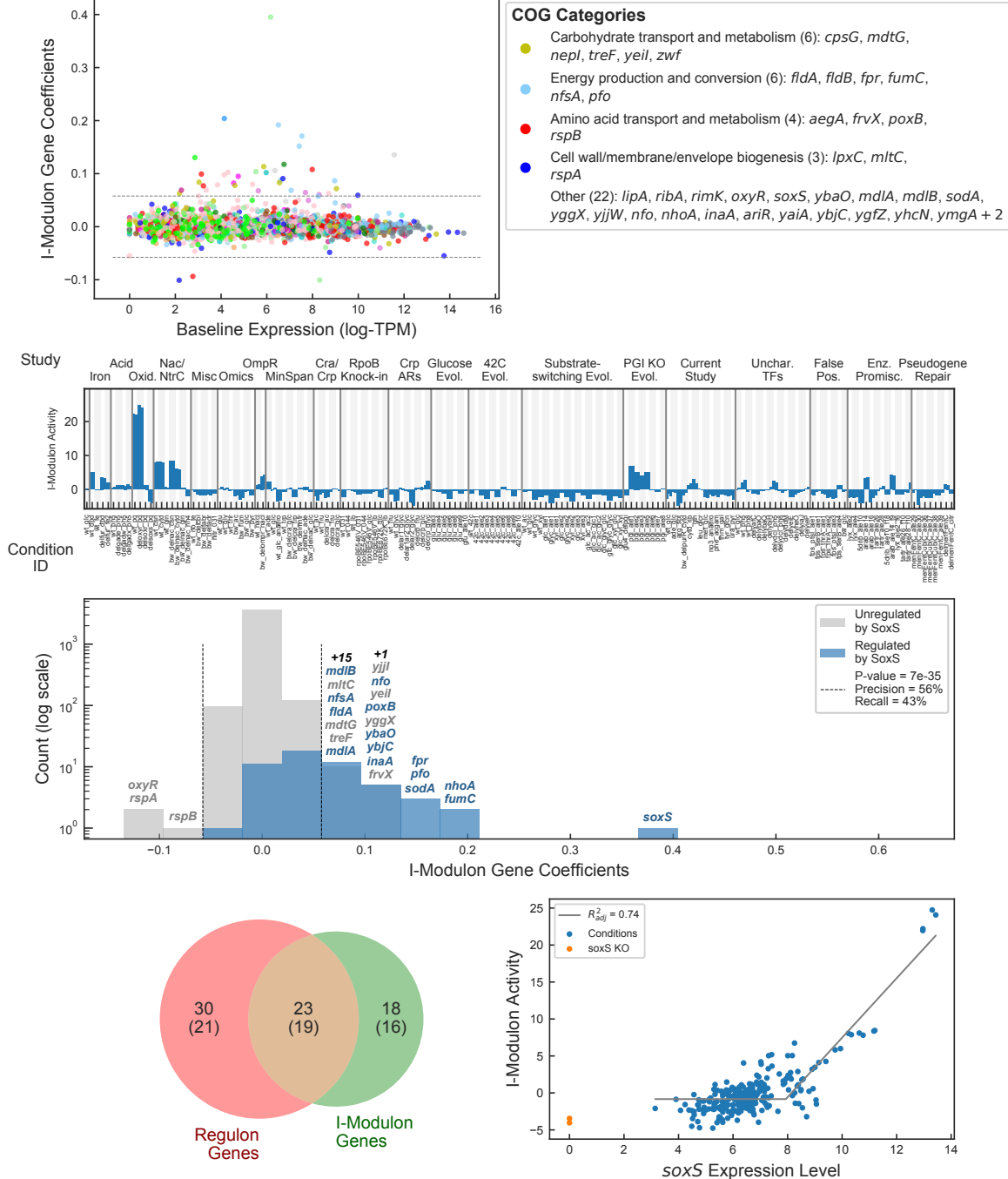
Biological Function: General stress response



SoxS I-Modulon

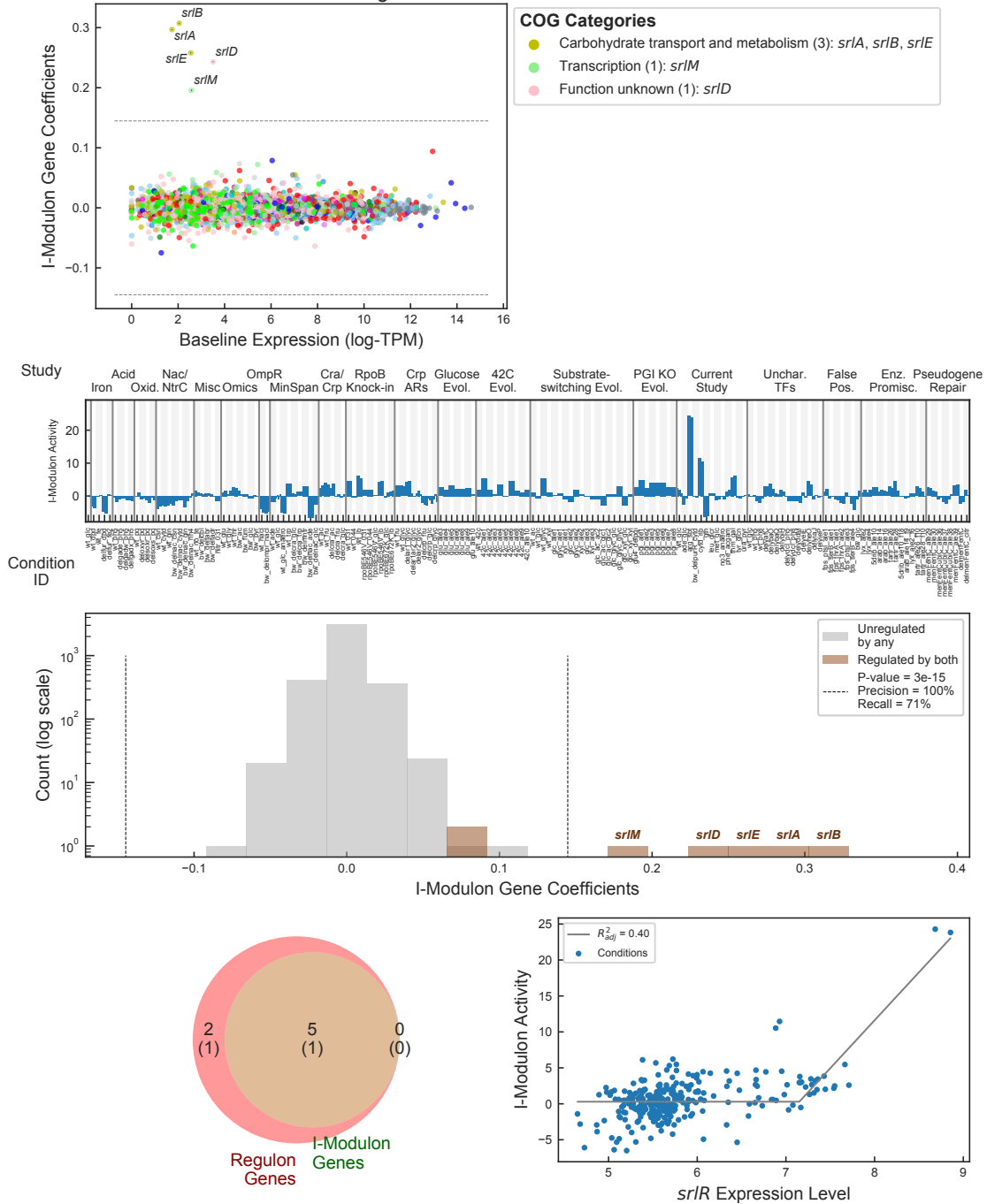
Regulated by: SoxS

Biological Function: Oxidative stress response



SrlR + GutM I-Modulon

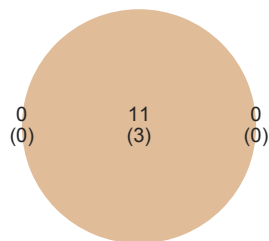
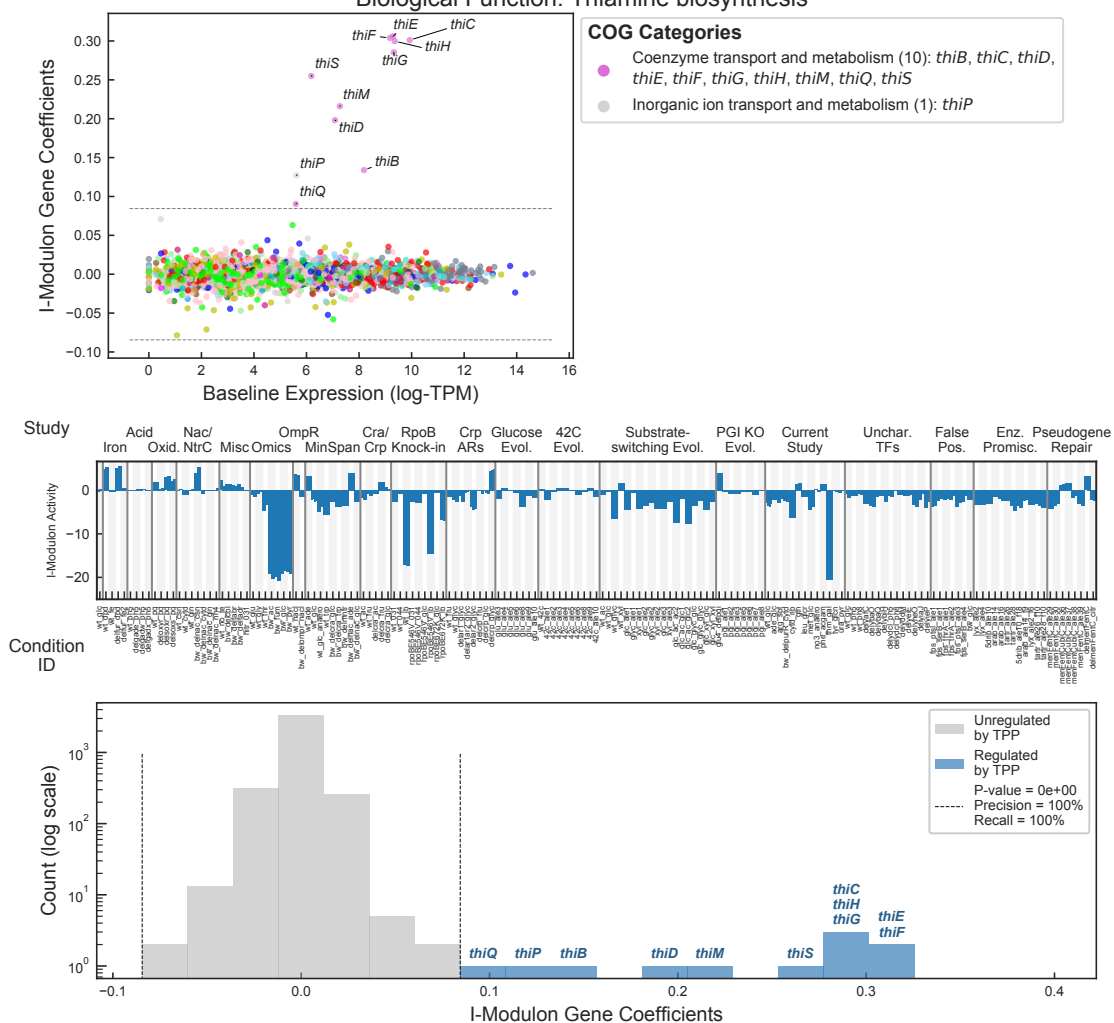
Regulated by: SrlR and GutM
Biological Function: Sorbitol catabolism



Thiamine I-Modulon

Regulated by: Thiamine

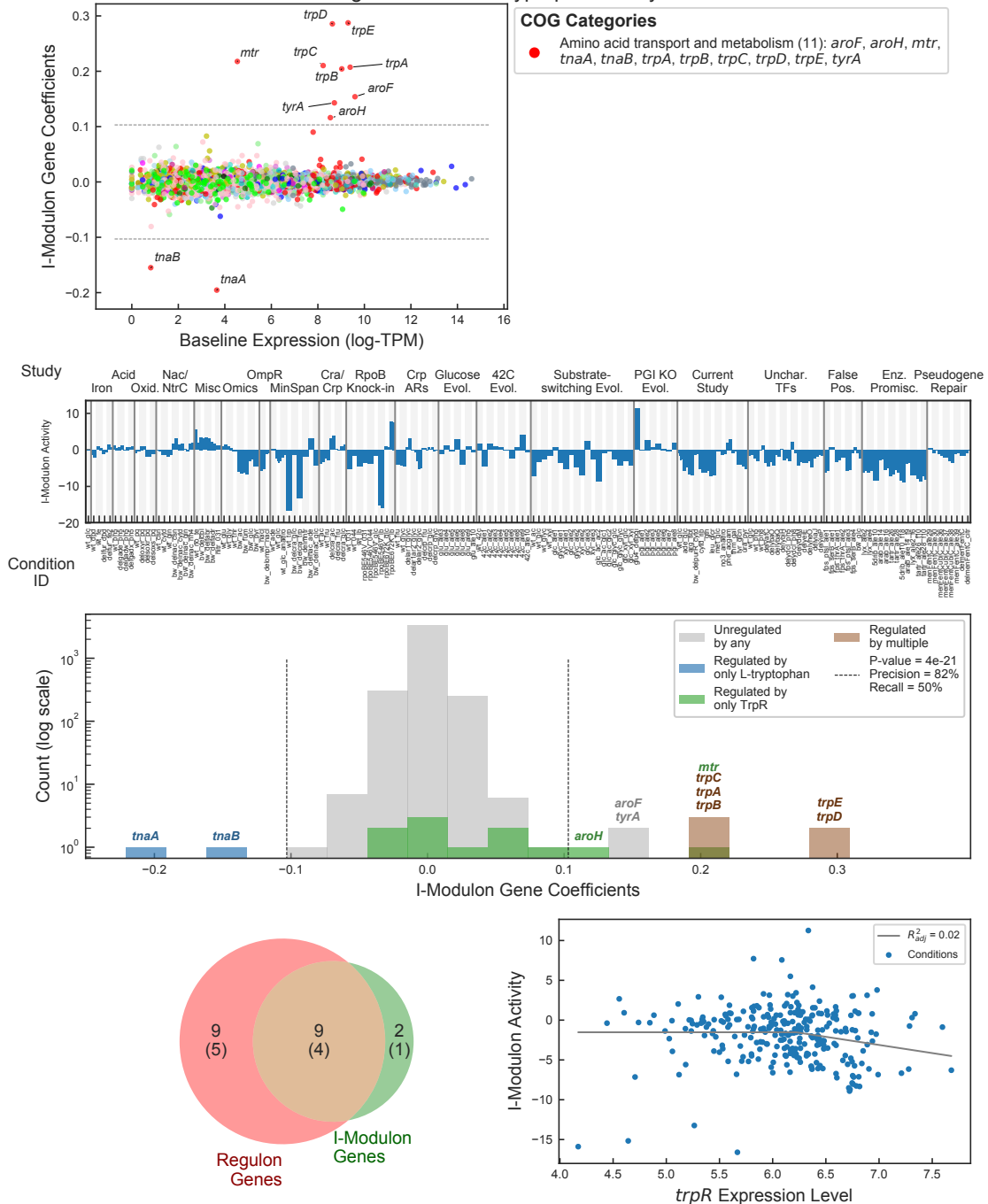
Biological Function: Thiamine biosynthesis



RegulonI-Modulon
GenesGenes

Tryptophan I-Modulon

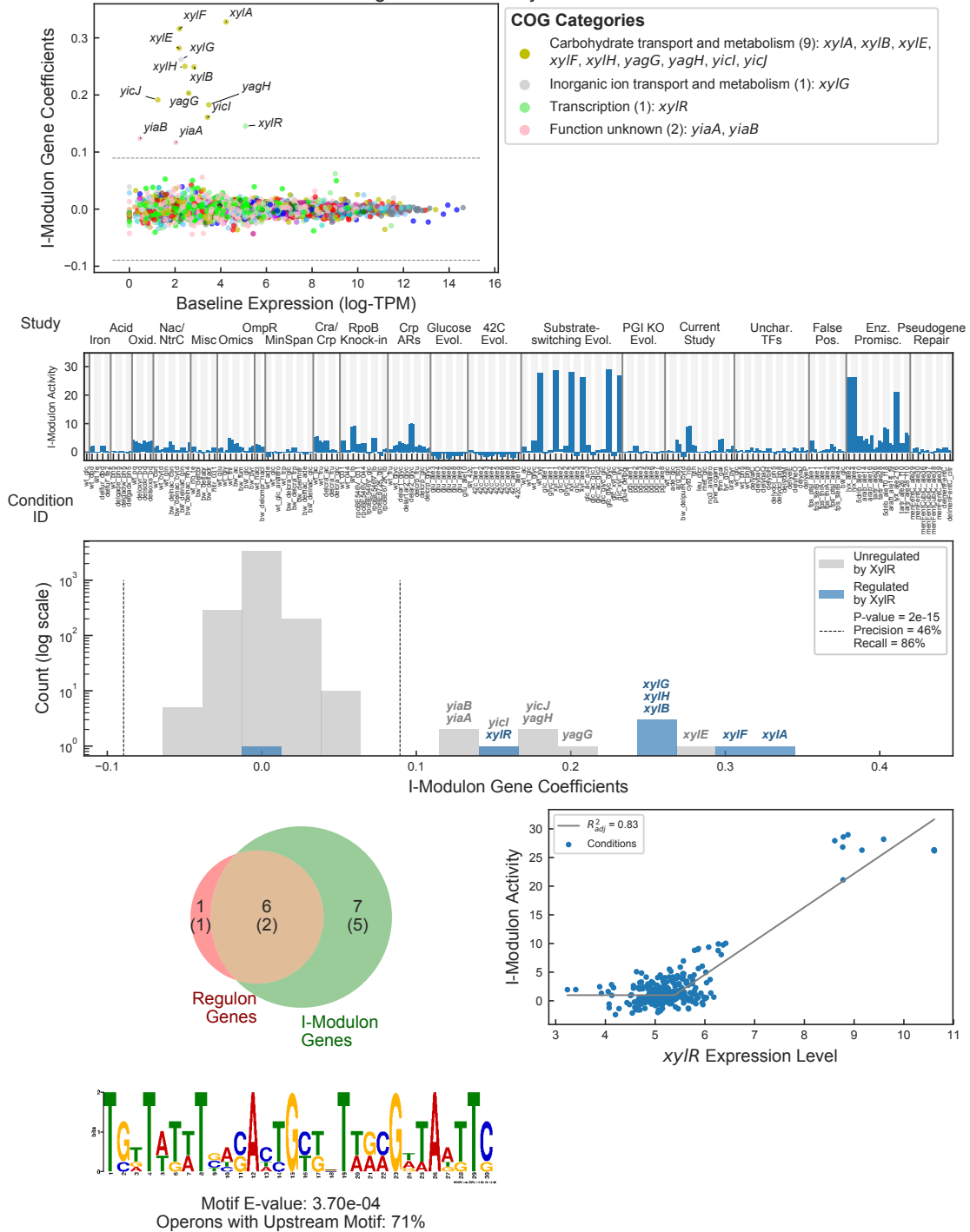
Regulated by: TrpR or trp-tRNA attenuation or Tryptophan attenuation
Biological Function: Tryptophan Biosynthesis



XylR I-Modulon

Regulated by: XylR

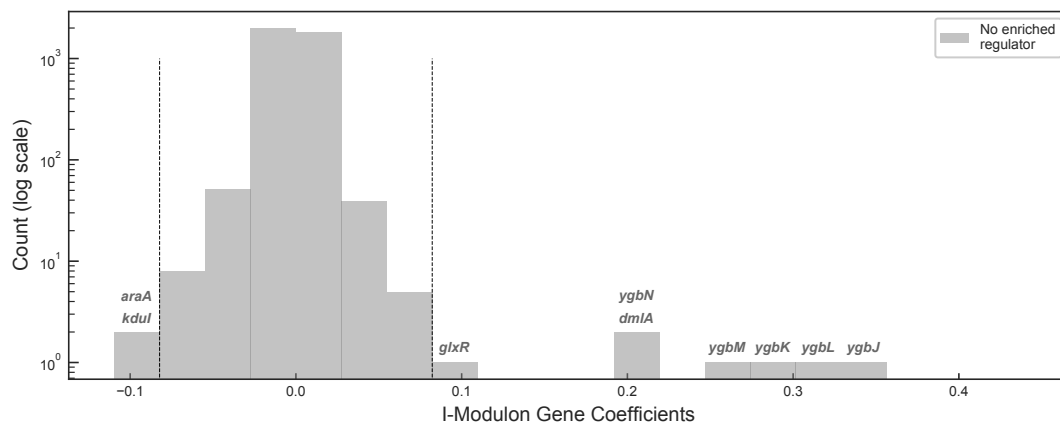
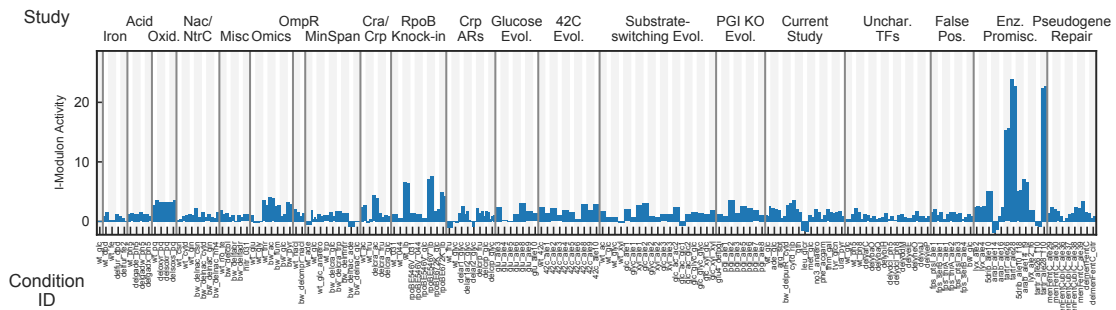
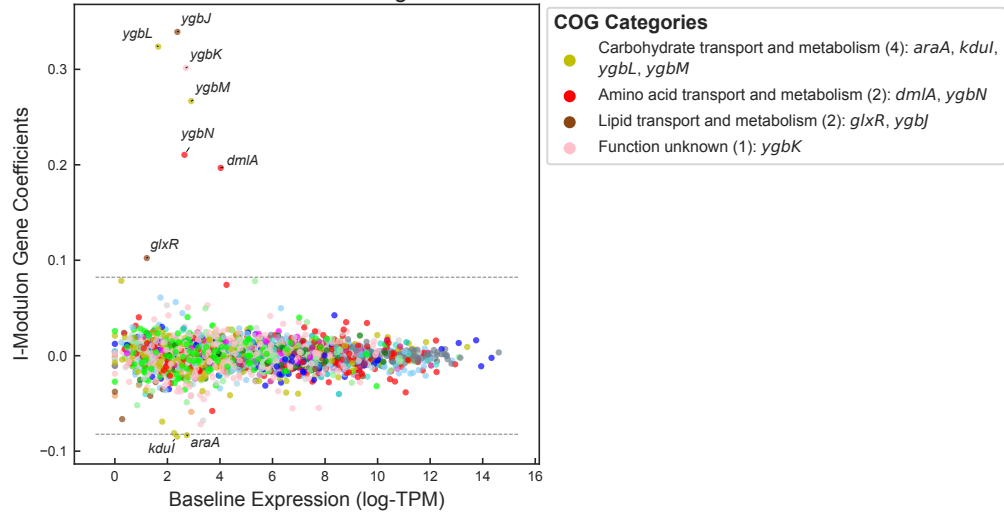
Biological Function: Xylose catabolism



YgbI I-Modulon

Regulated by: YgbI

Biological Function: Unknown Function



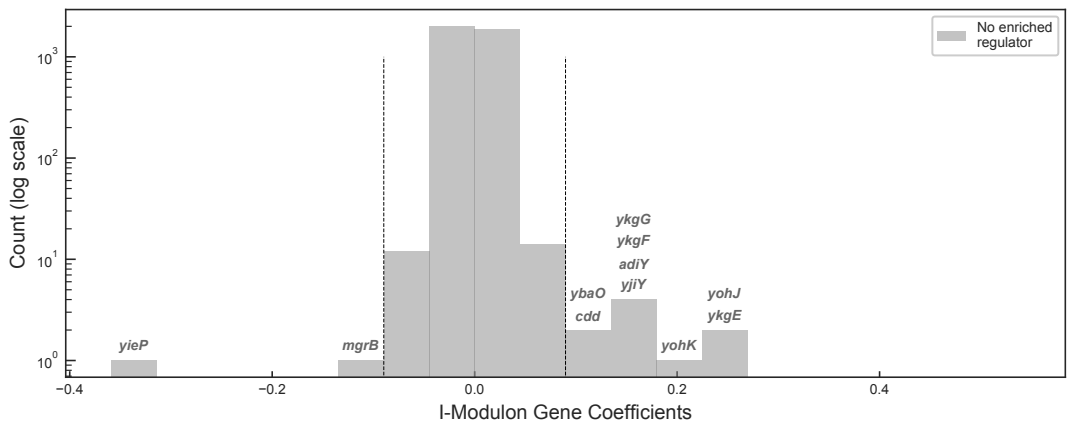
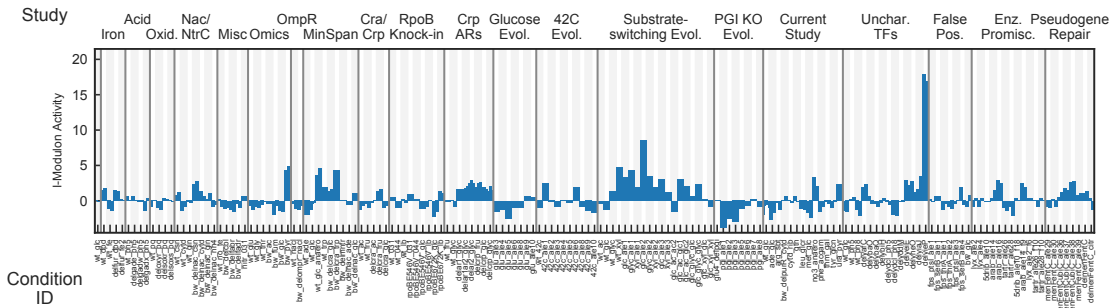
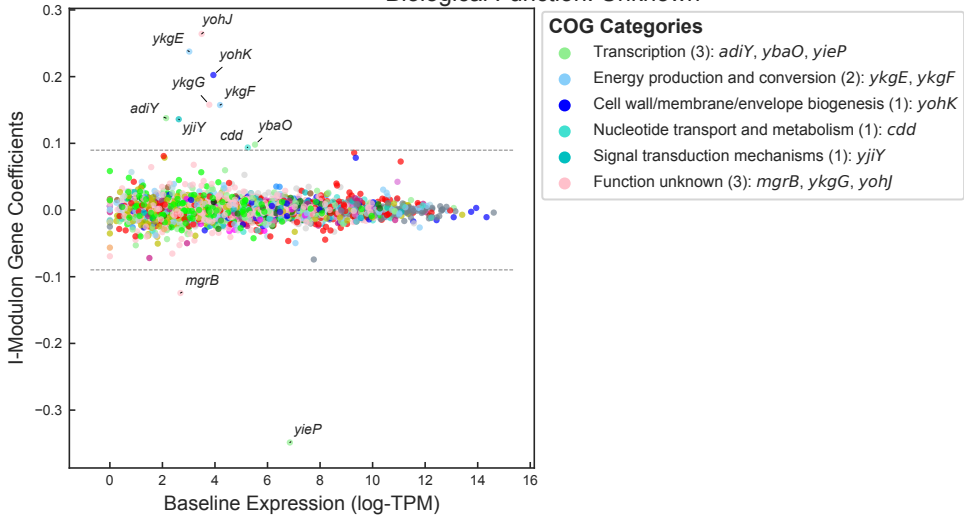
Regulated by: YiaJ
Biological Function: Ascorbate utilization



YieP I-Modulon

Regulated by: YieP

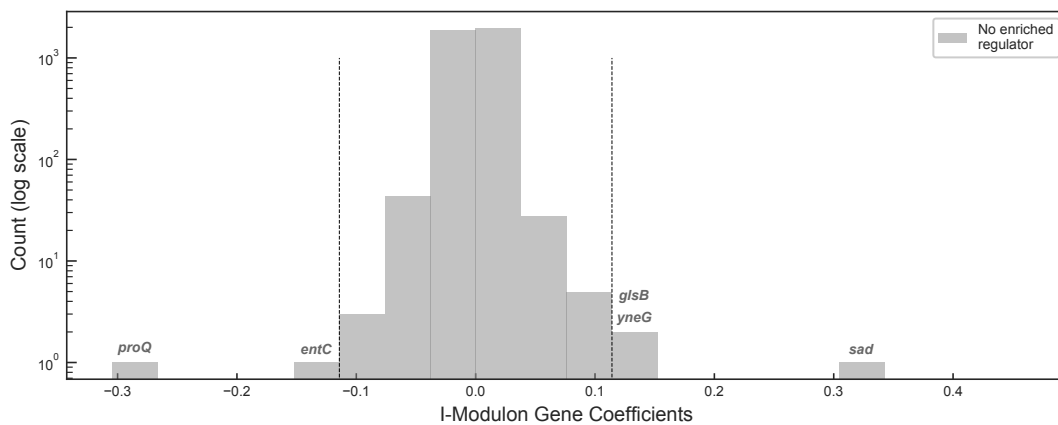
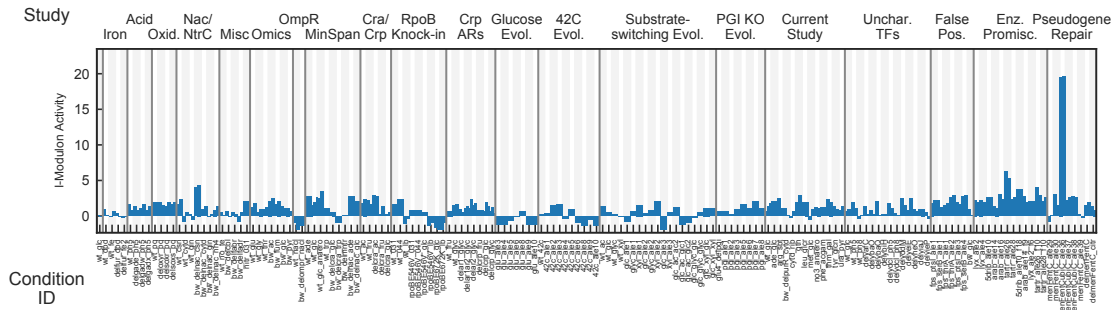
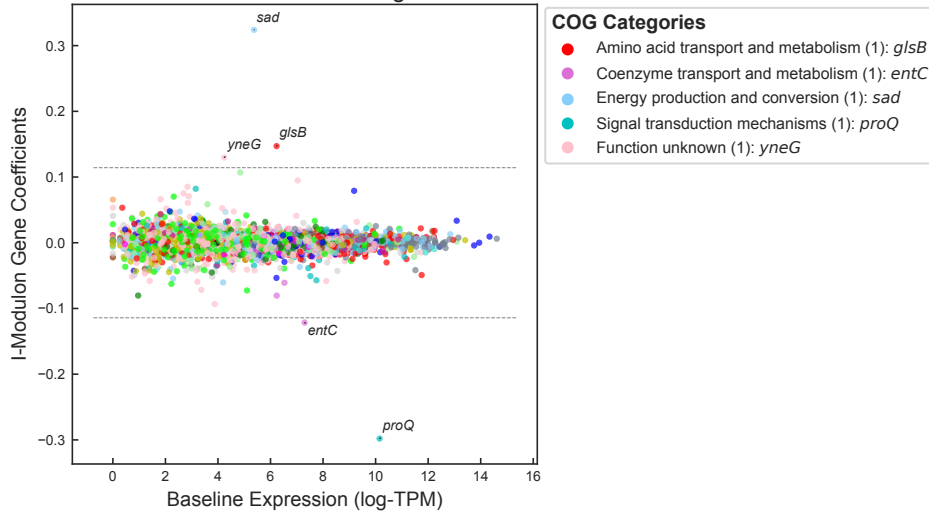
Biological Function: Unknown



YneJ I-Modulon

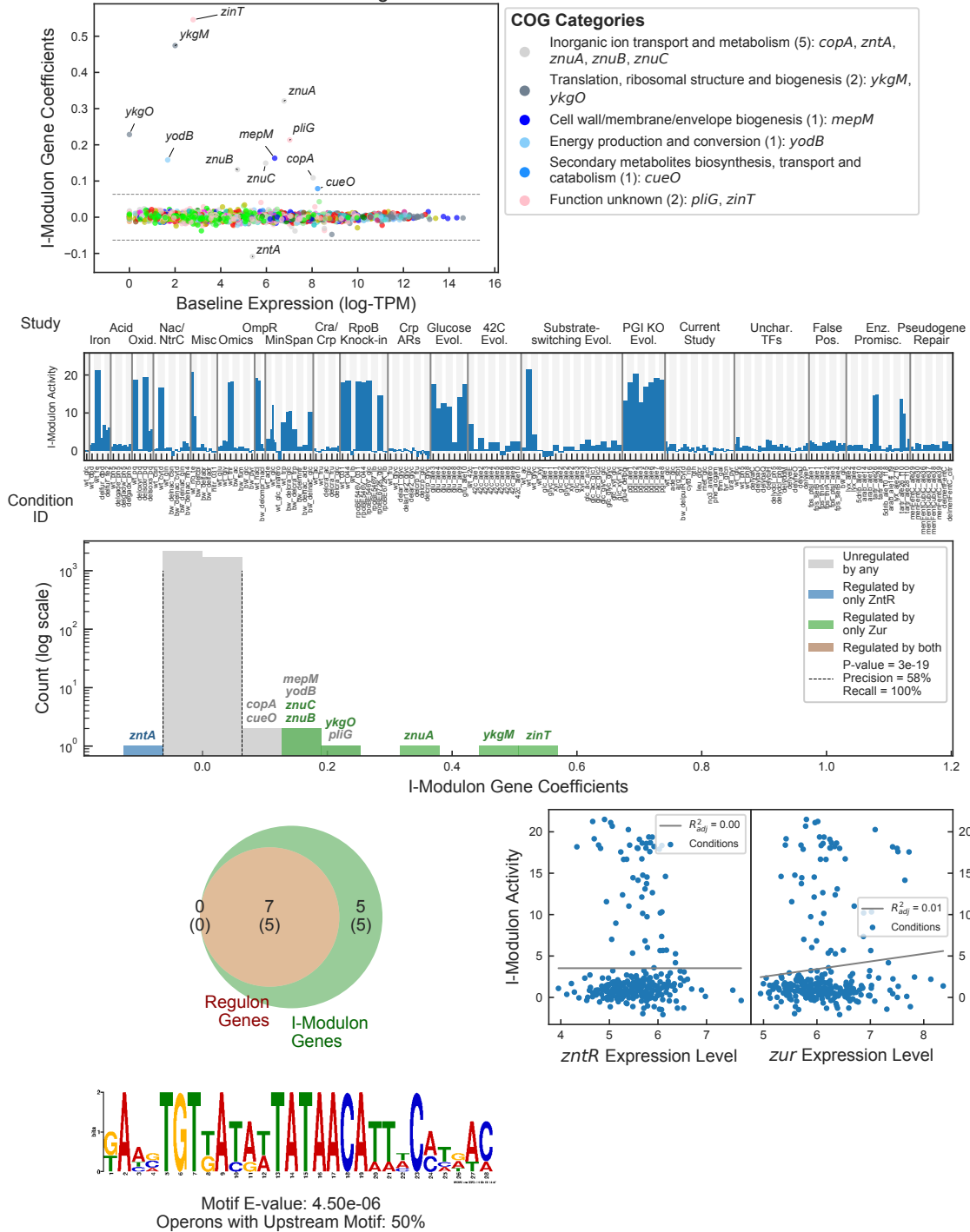
Regulated by: YneJ

Biological Function: Unknown Function



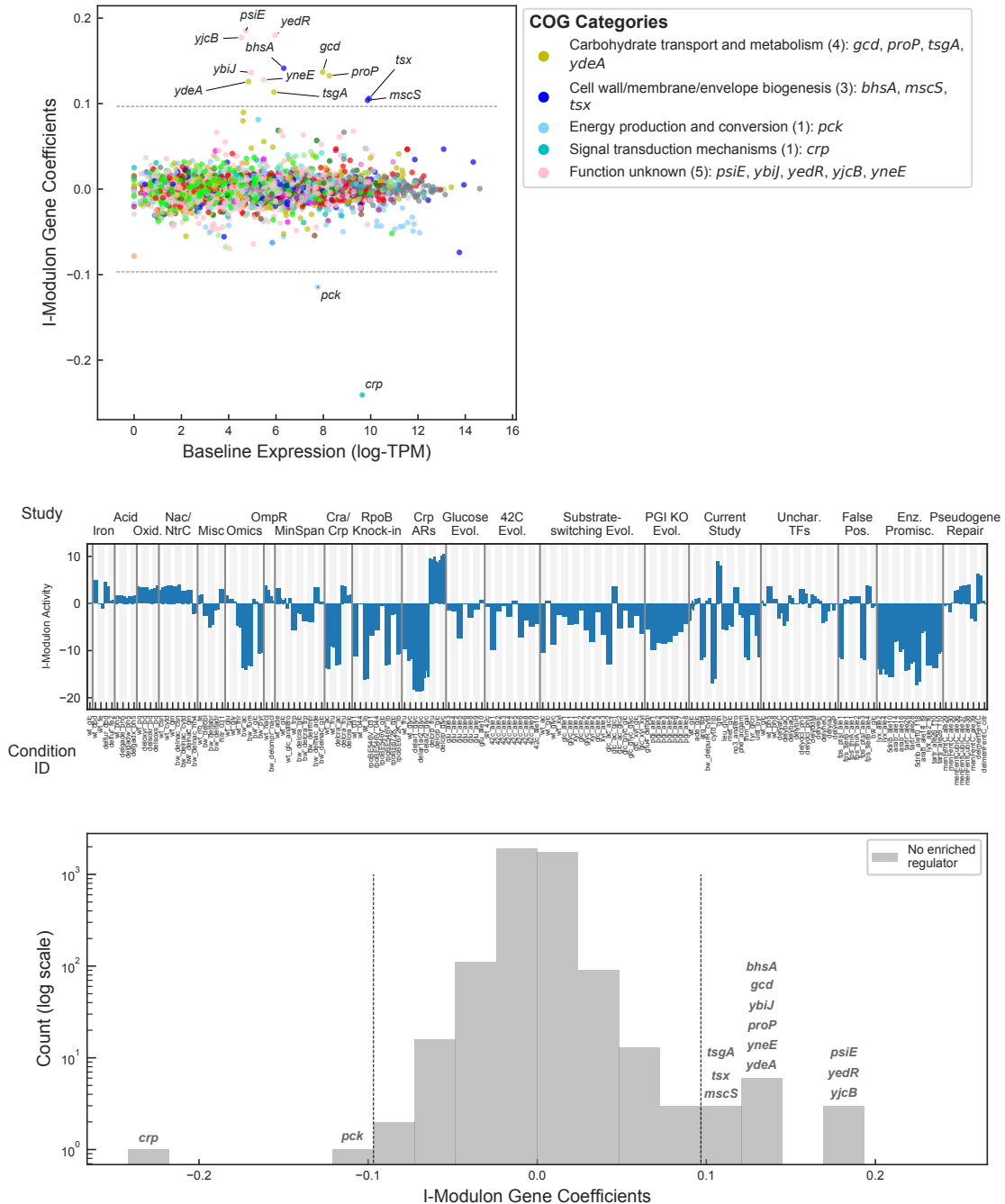
Zinc I-Modulon

Regulated by: ZntR or Zur
Biological Function: Zinc homeostasis



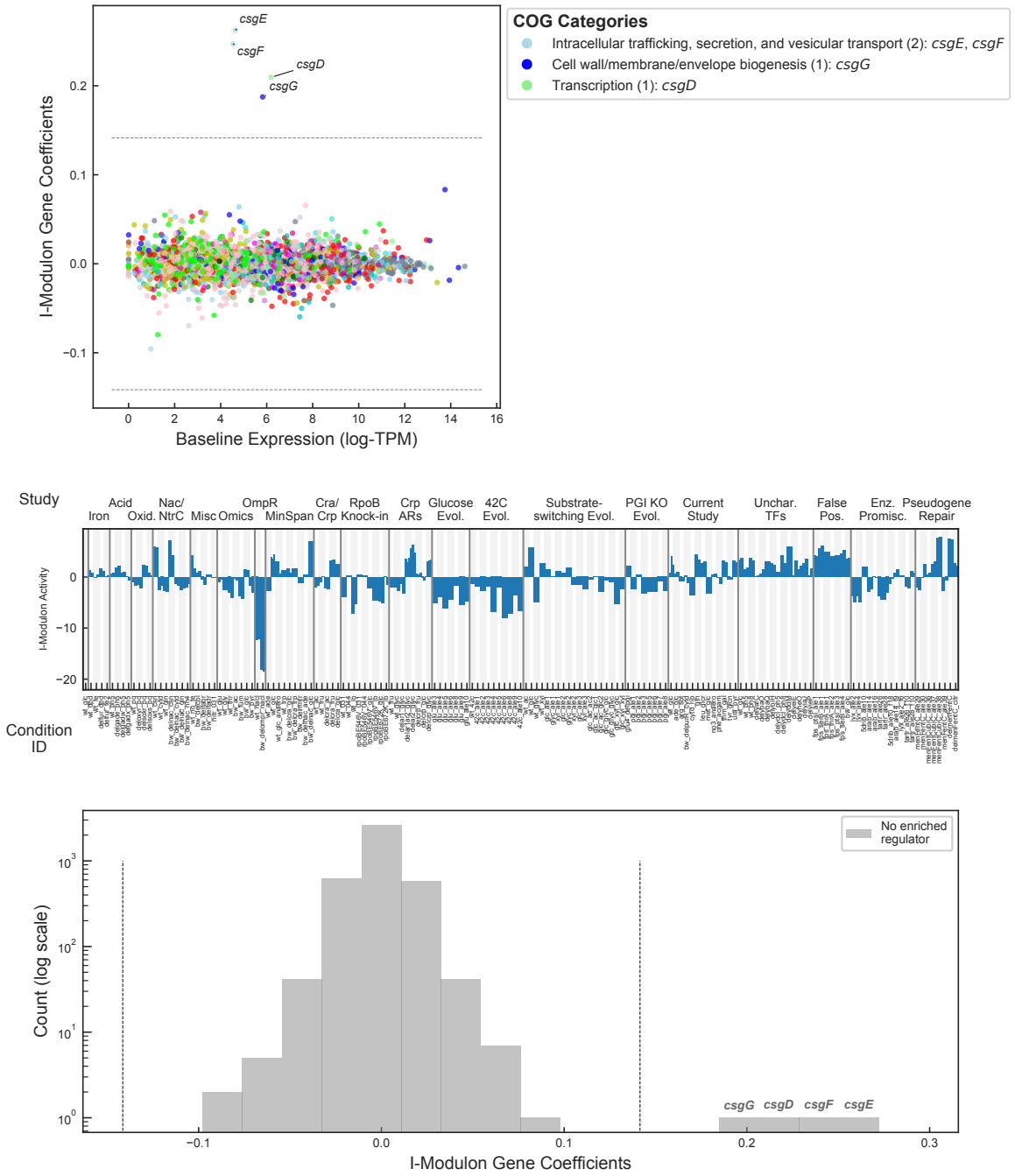
crp – KO I-Modulon

Biological Function: Accounts for crp knock-out



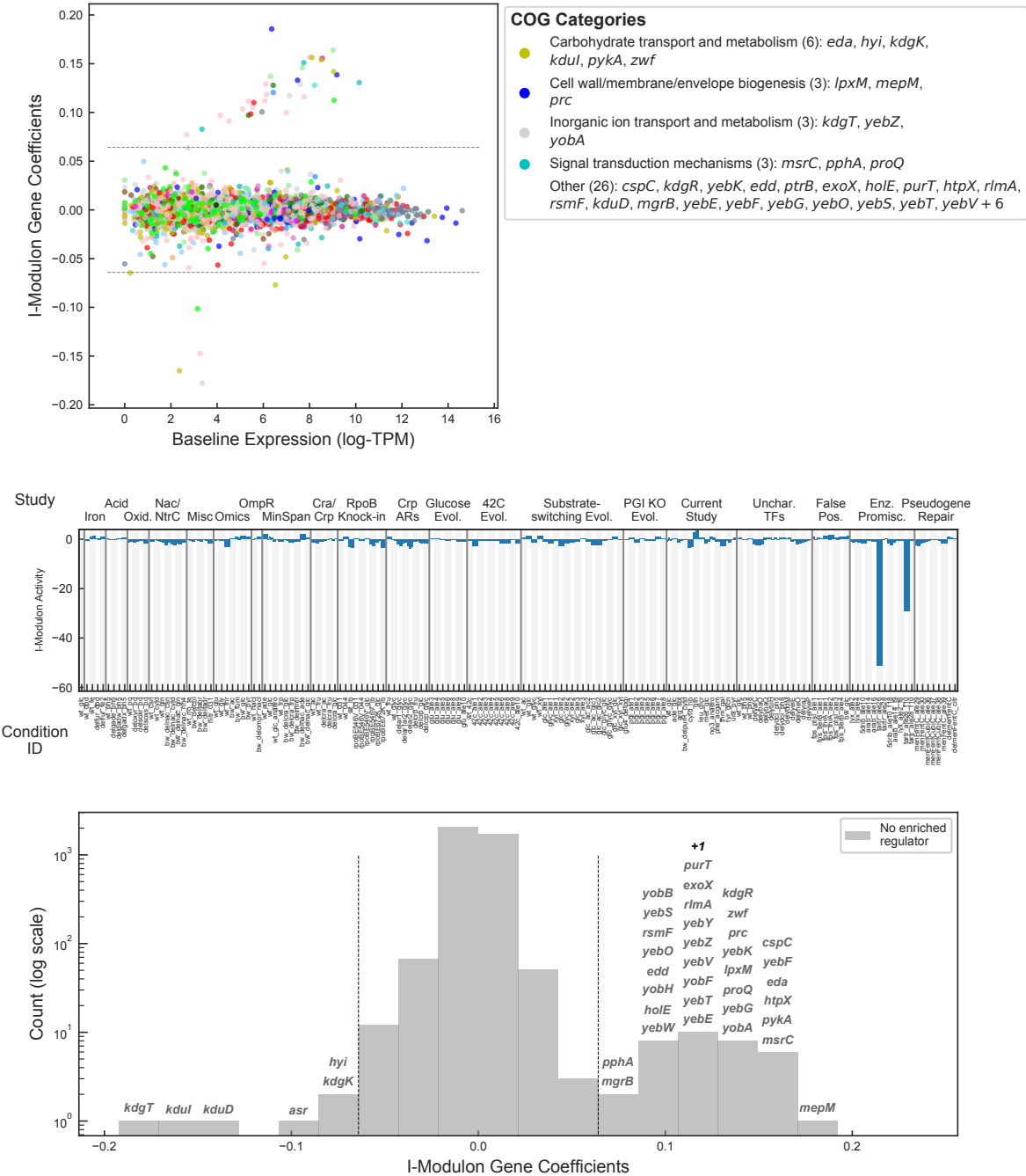
curli I-Modulon

Biological Function: Curli assembly



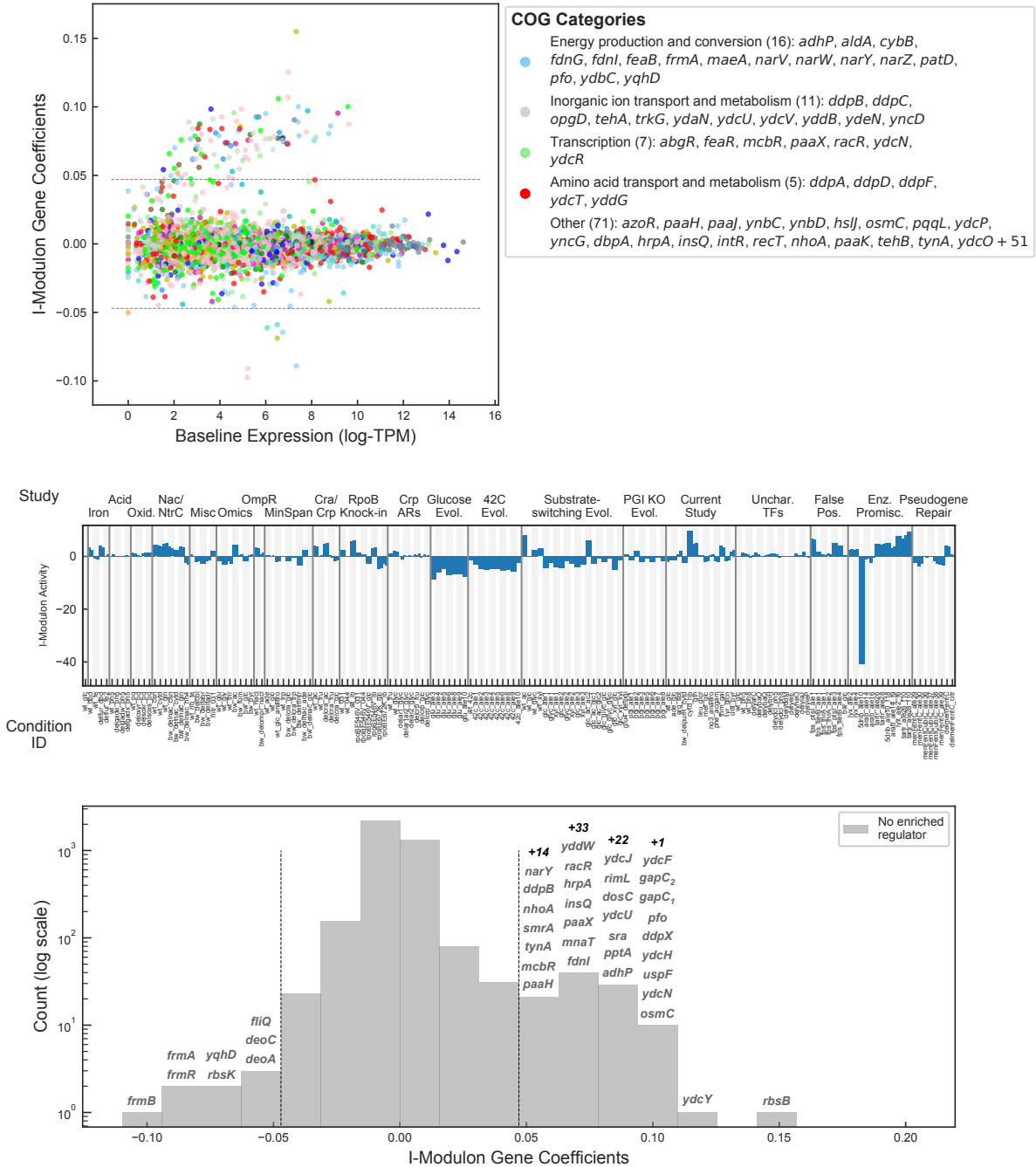
deletion – 1 I-Modulon

Biological Function: Large deletion of 39 genes during evolution



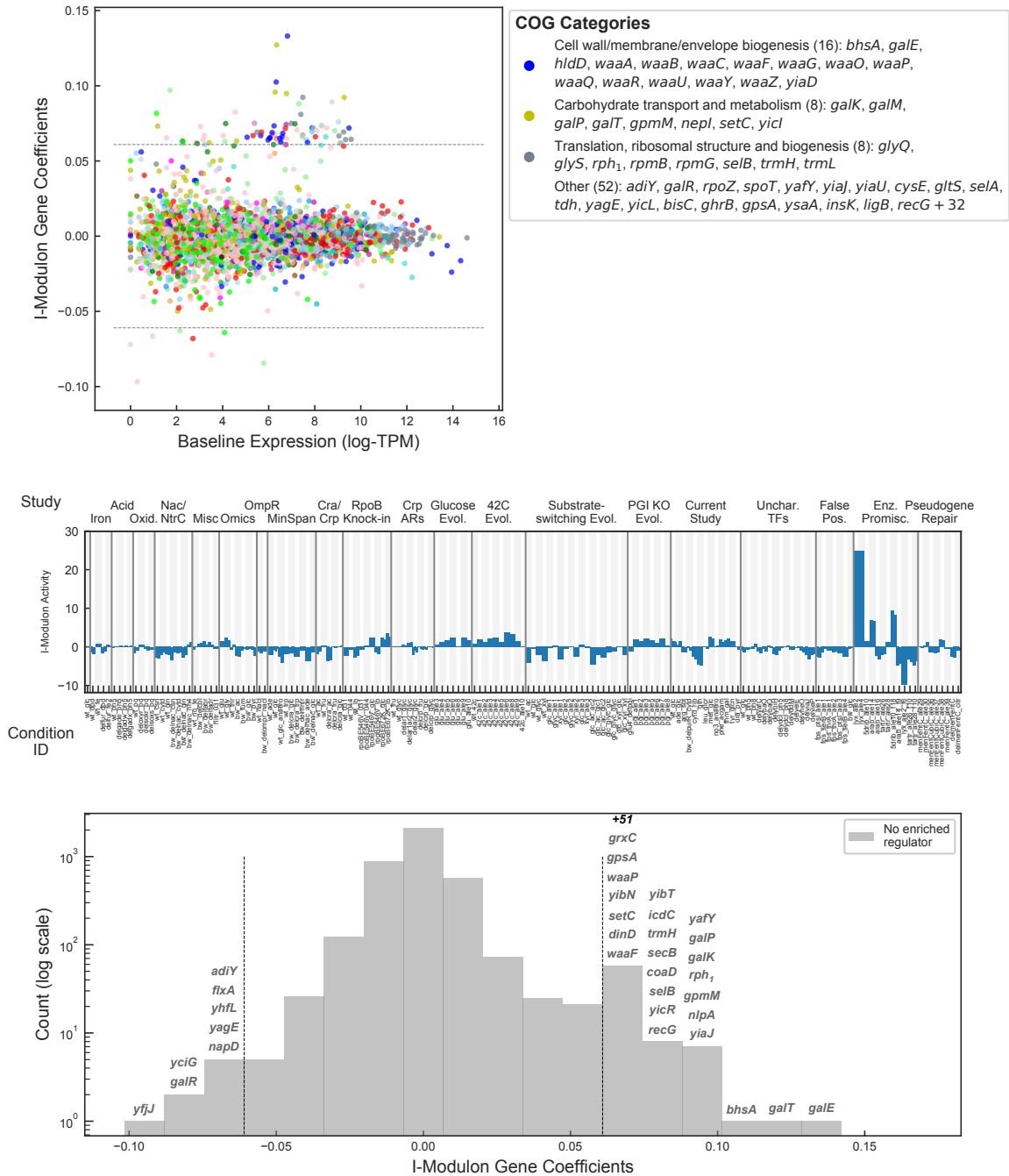
deletion – 2 I-Modulon

Biological Function: Large deletion of 171 genes during evolution



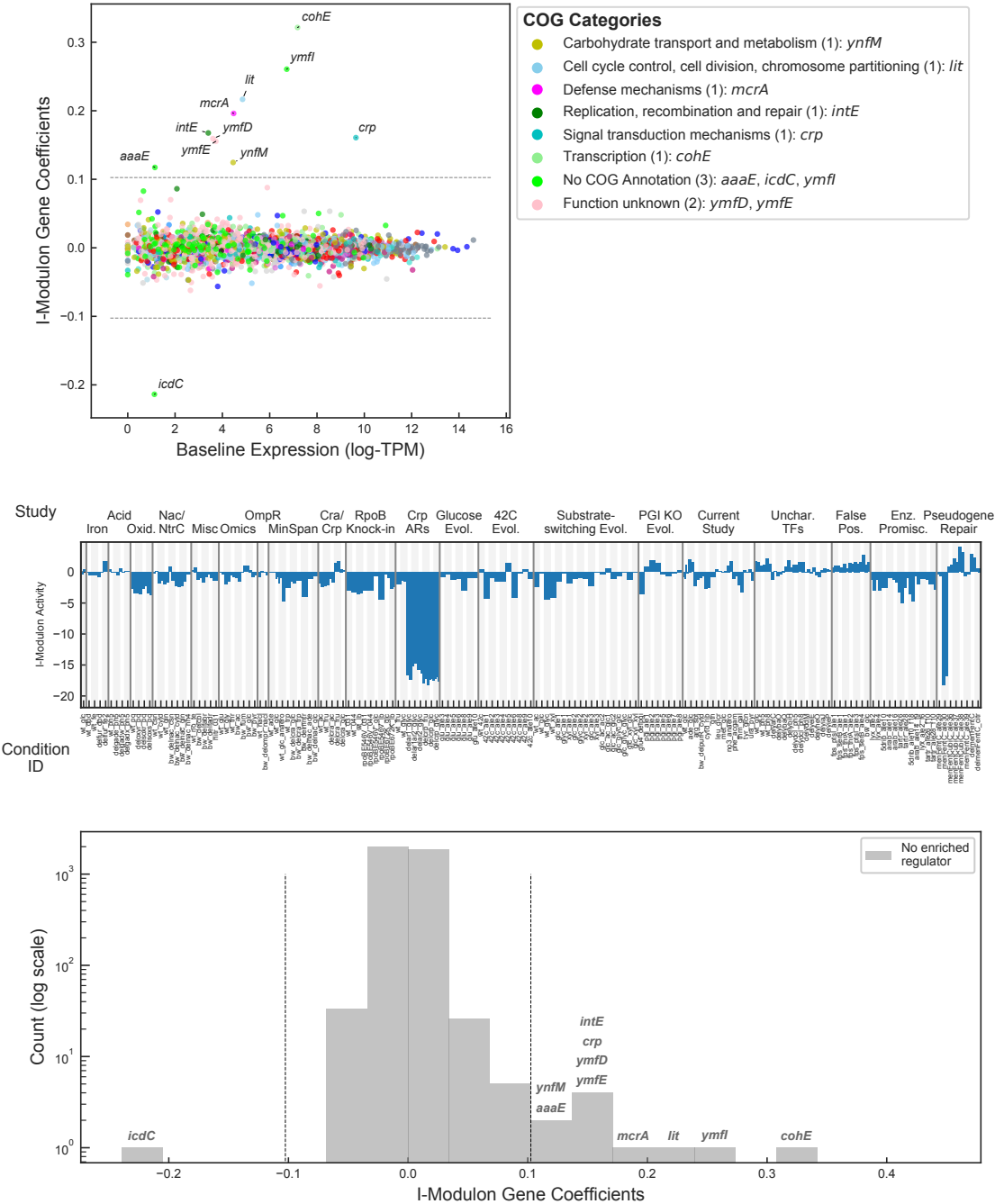
duplication – 1 I-Modulon

Biological Function: Large duplication of 129 genes during evolution



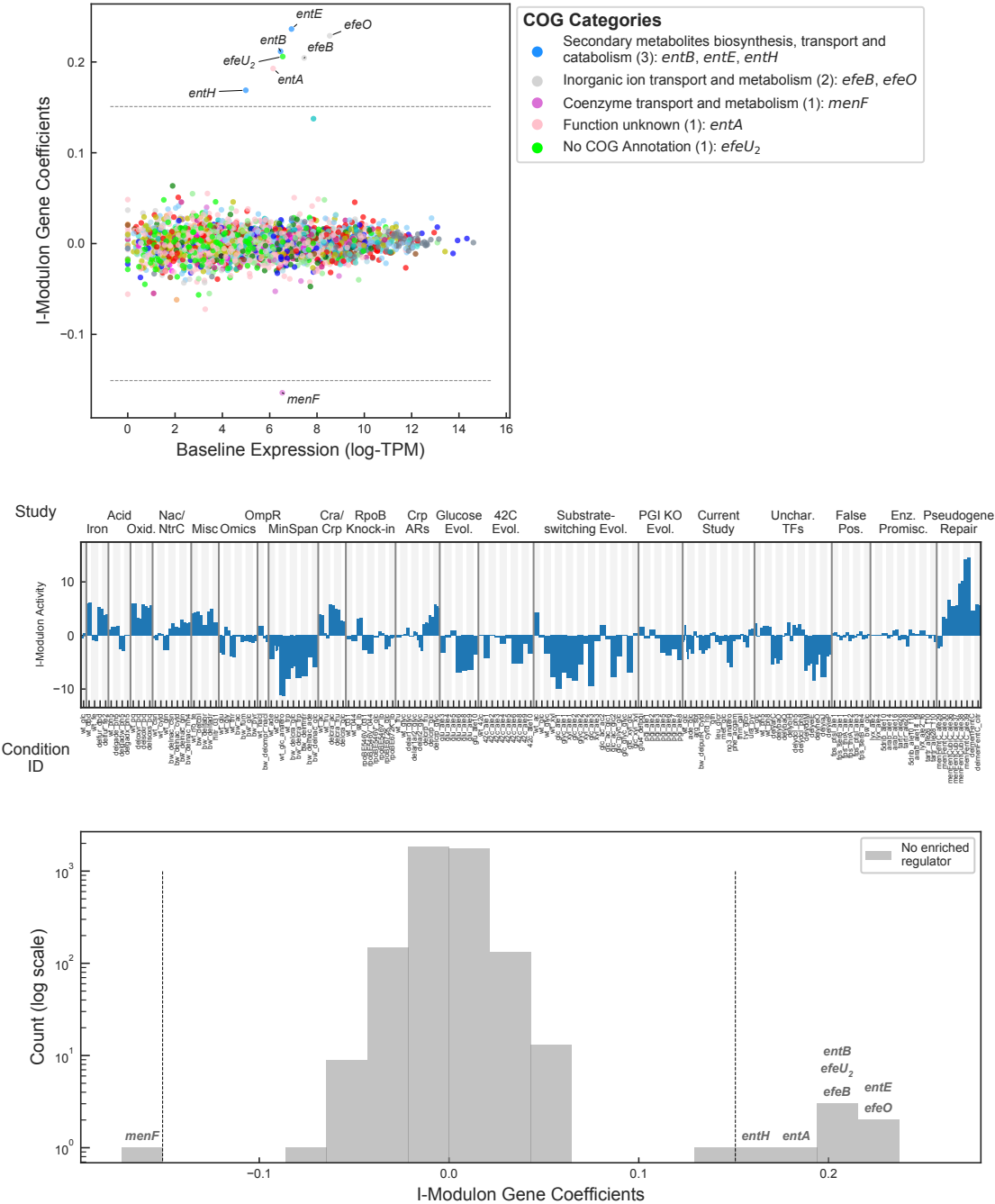
e14 – deletion I-Modulon

Biological Function: Removal of e14 prophage



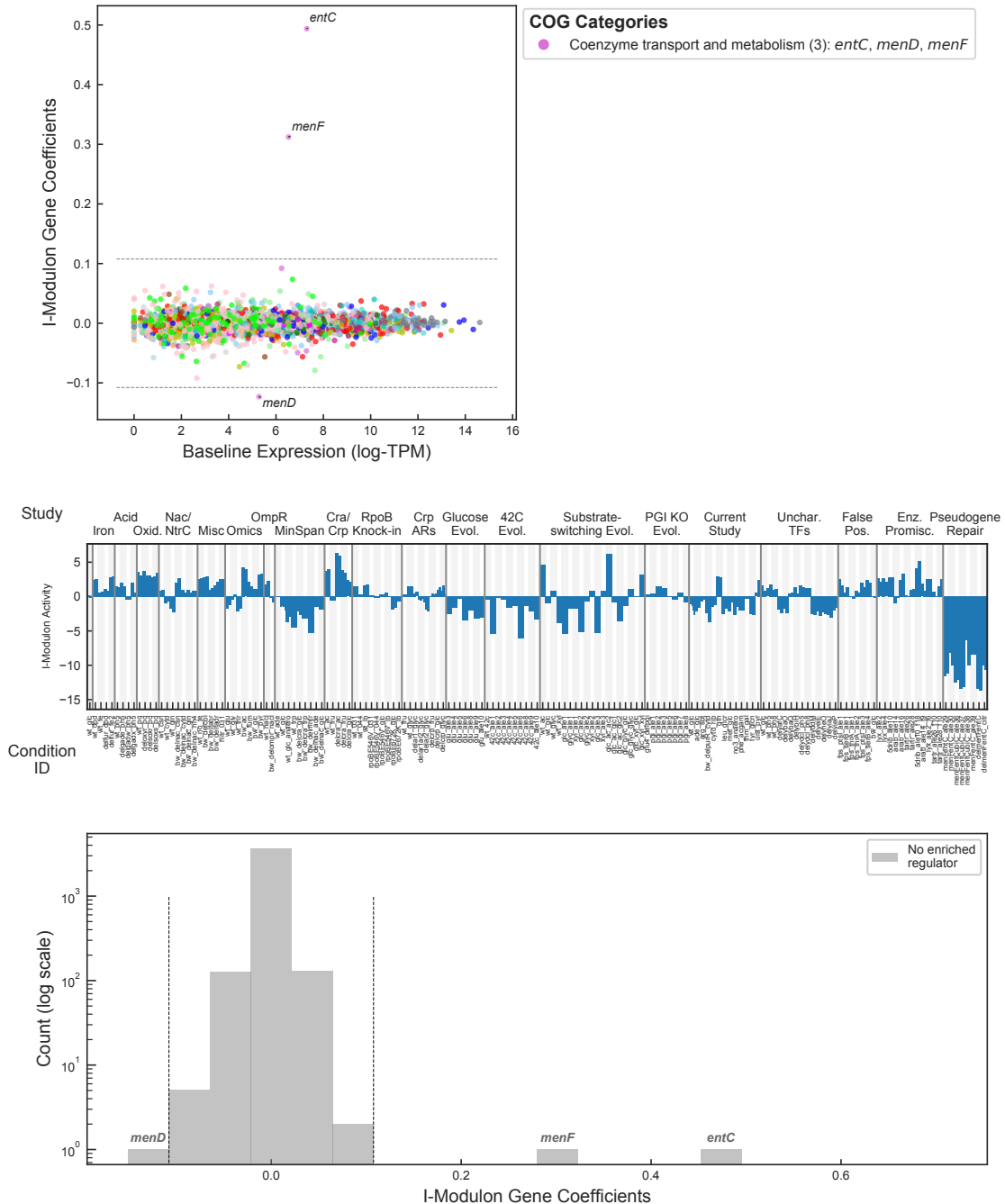
efeU – repair I-Modulon

Biological Function: Accounts for repair and expression of efeU operon



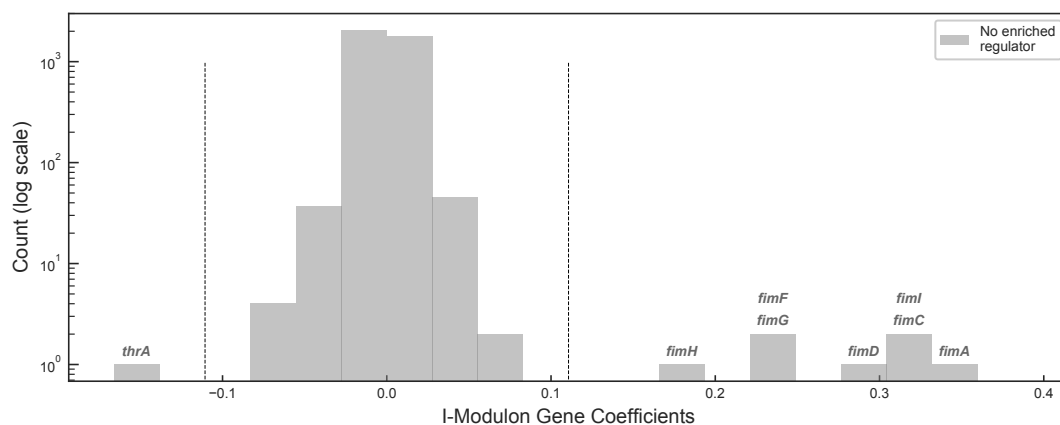
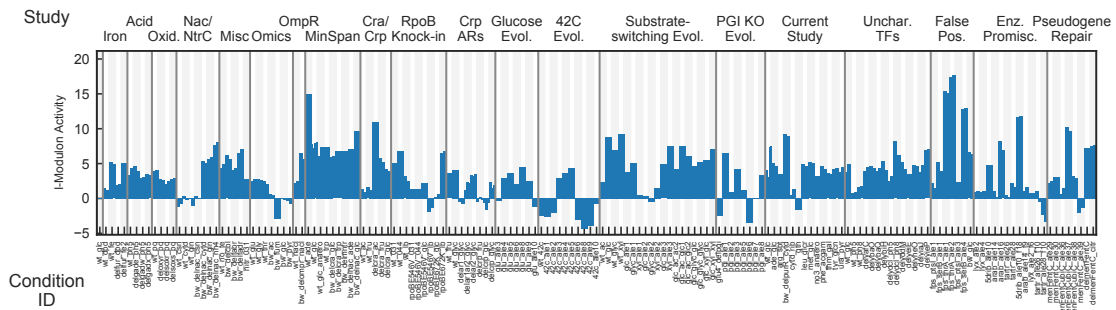
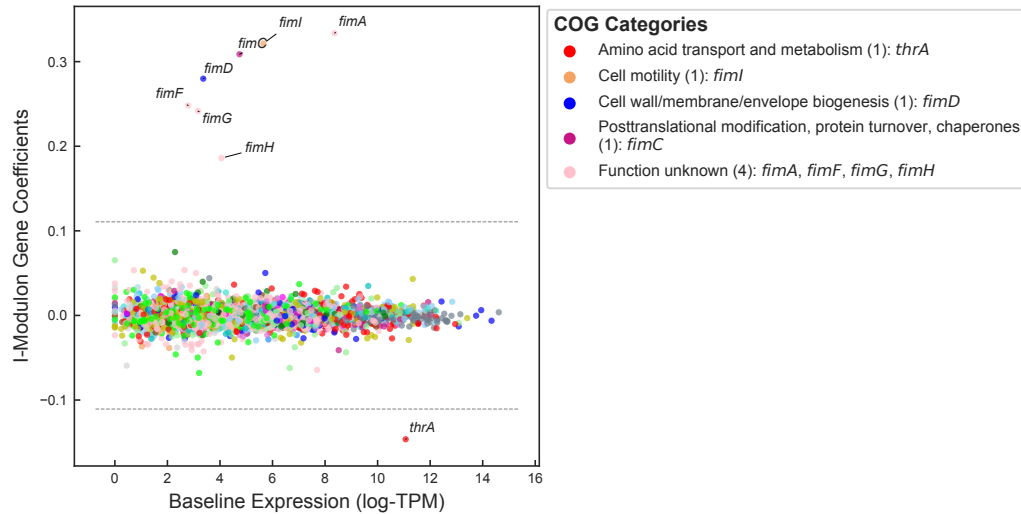
entC – menF – KO I-Modulon

Biological Function: Accounts for entC and menF knock-outs



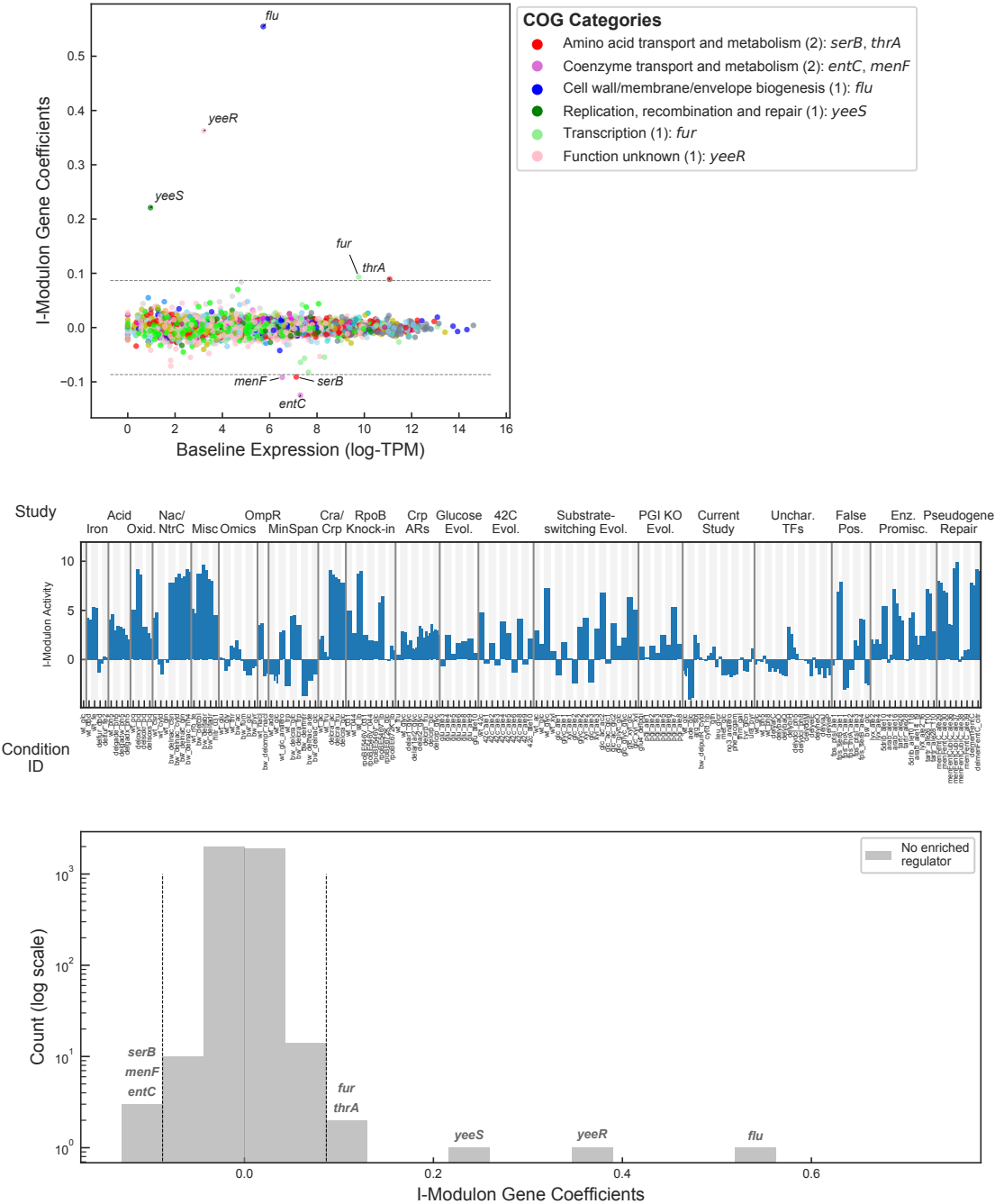
fimbriae I-Modulon

Biological Function: Fimbriae assembly



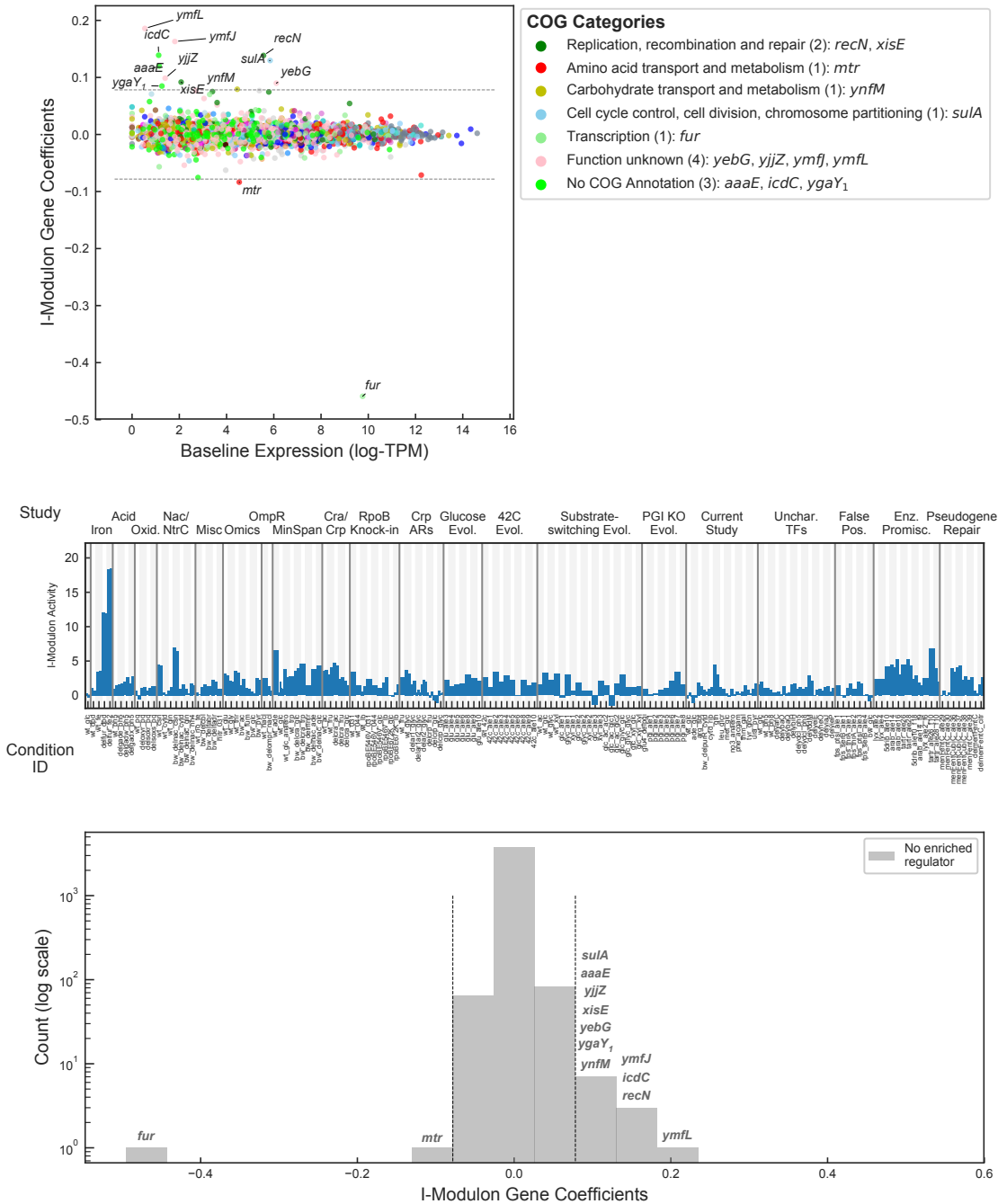
flu – yeeRS I-Modulon

Biological Function: Genes in CP4-44 prophage



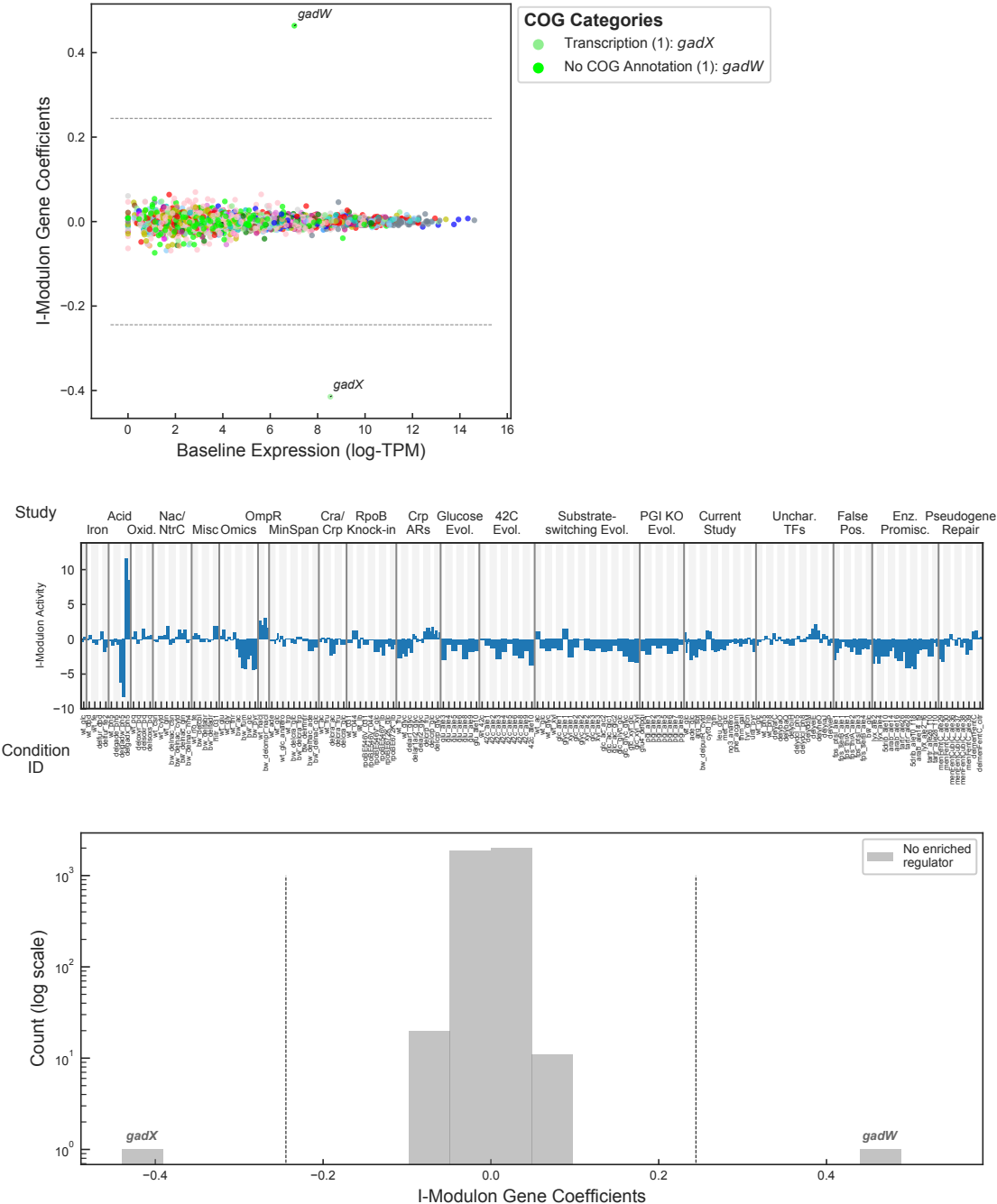
fur – KO I-Modulon

Biological Function: Accounts for fur knock-out



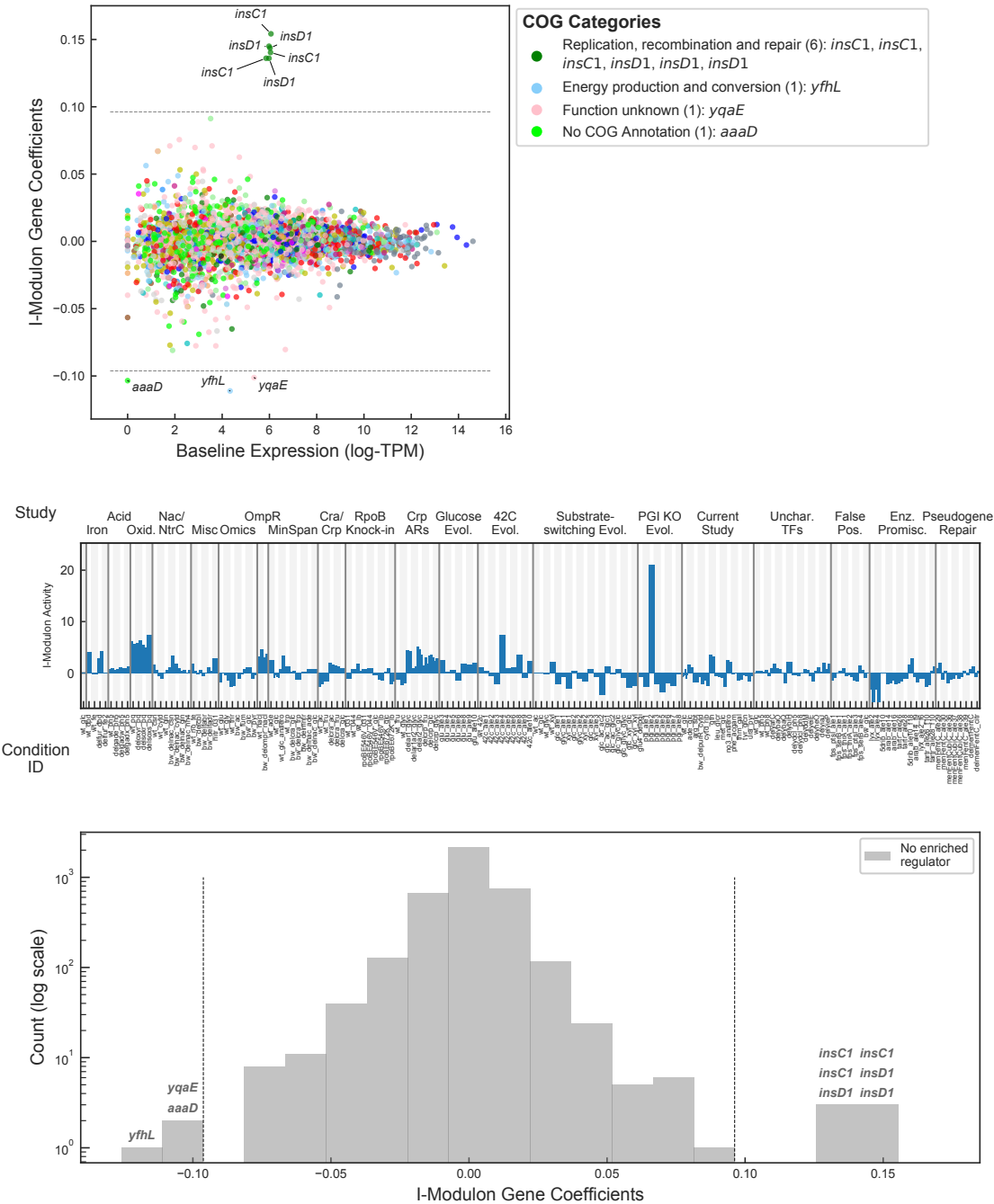
gadWX – KO I-Modulon

Biological Function: Accounts for gadW and gadX knock-outs



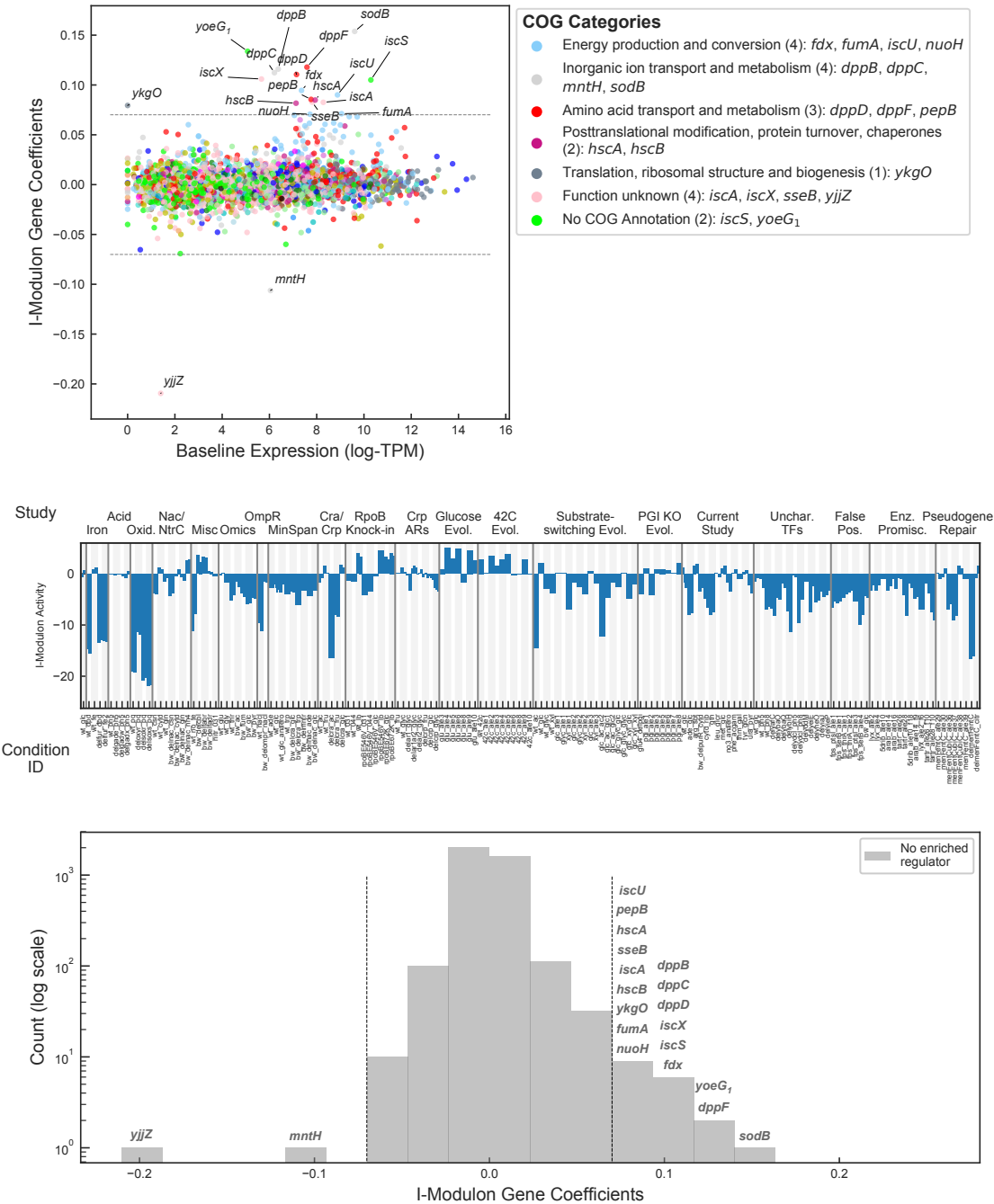
insertion I-Modulon

Biological Function: IS2 insertion element after laboratory evolution



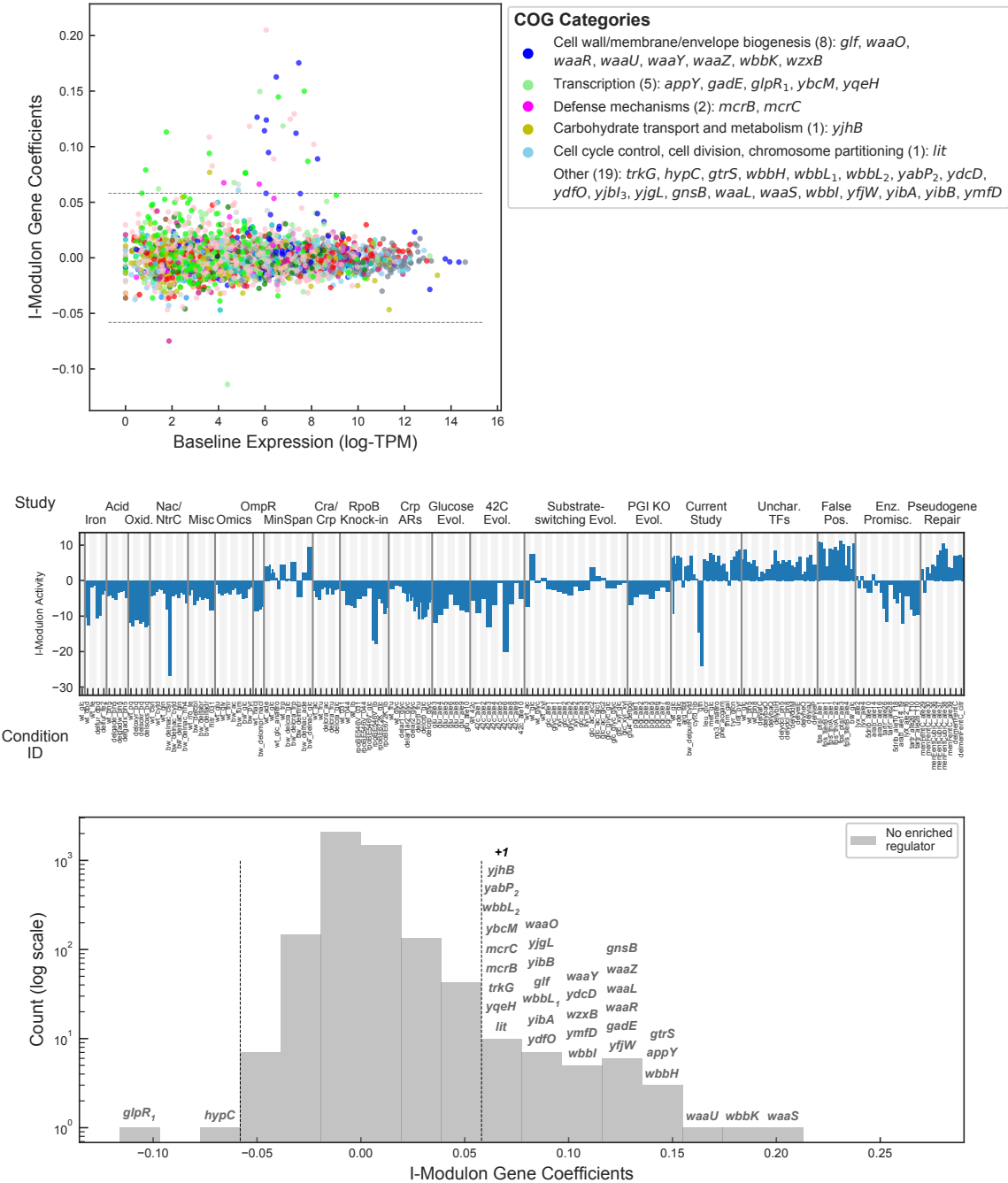
iron – related I-Modulon

Biological Function: Related to iron metabolism



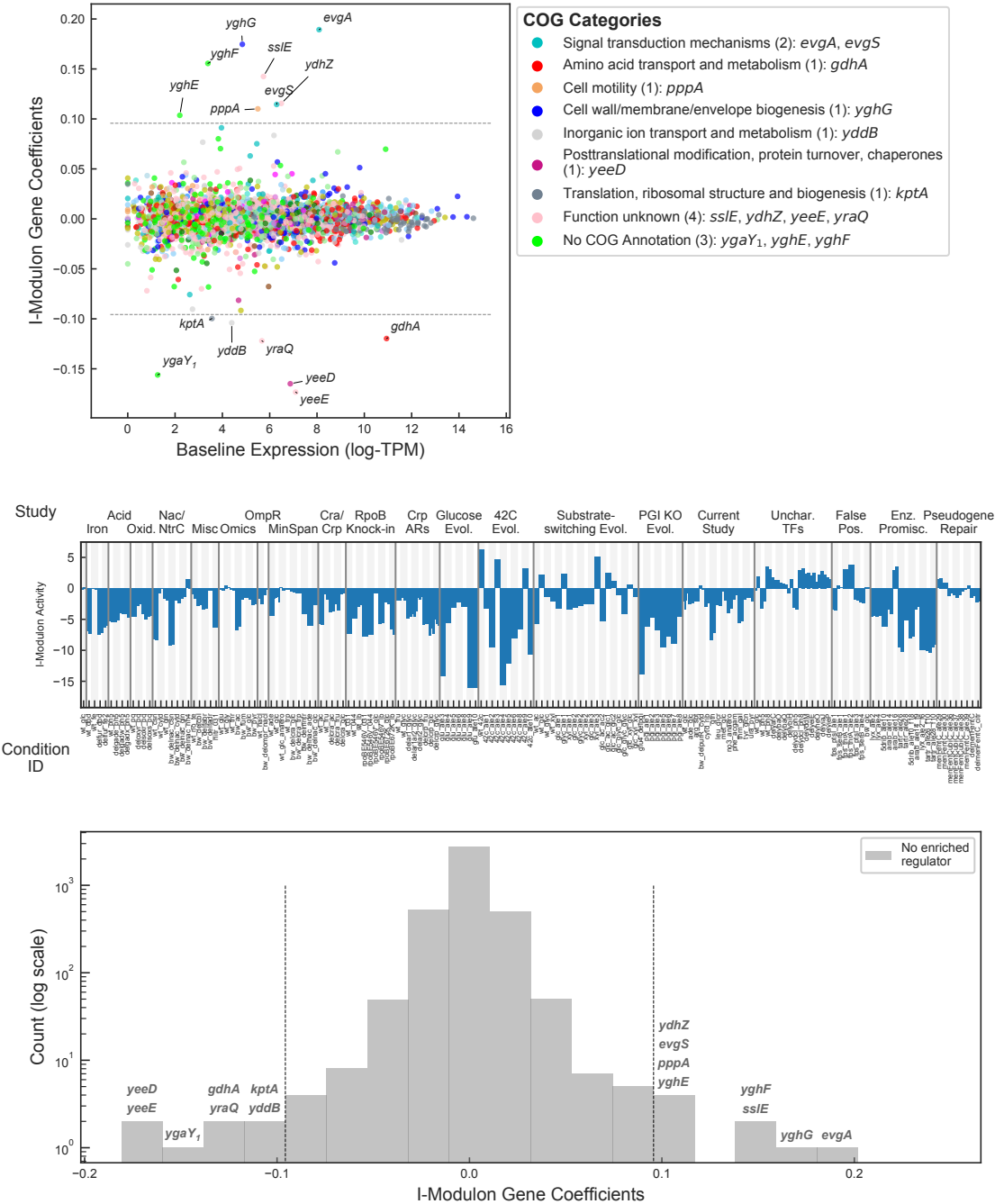
lipopolysaccharide I-Modulon

Biological Function: Lipopolysaccharide biosynthesis



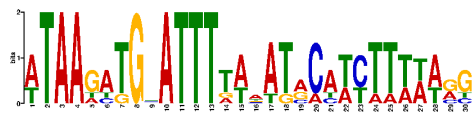
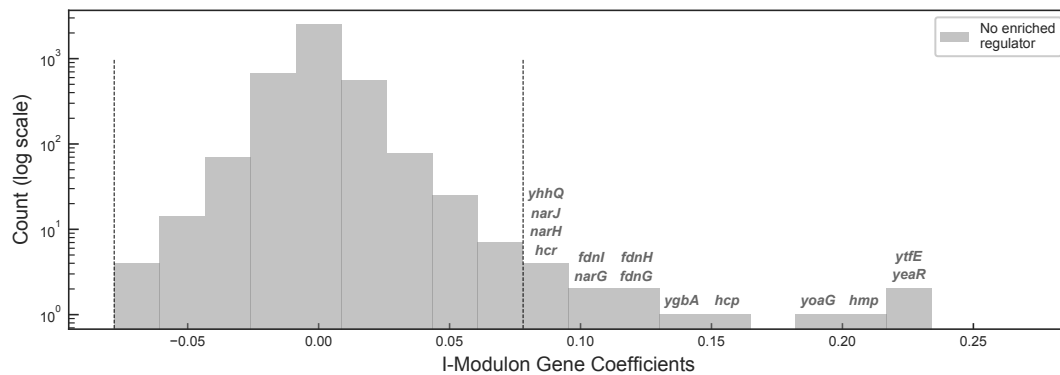
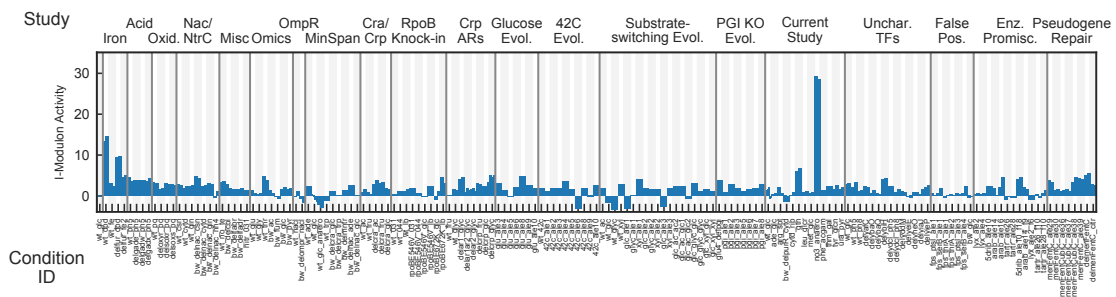
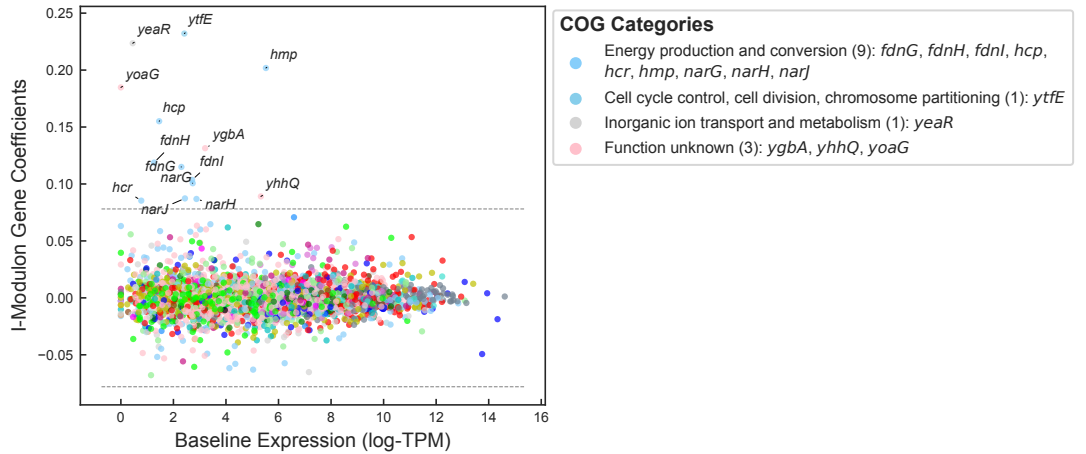
membrane I-Modulon

Biological Function: Enriched in membrane-bound proteins



nitrate – related I-Modulon

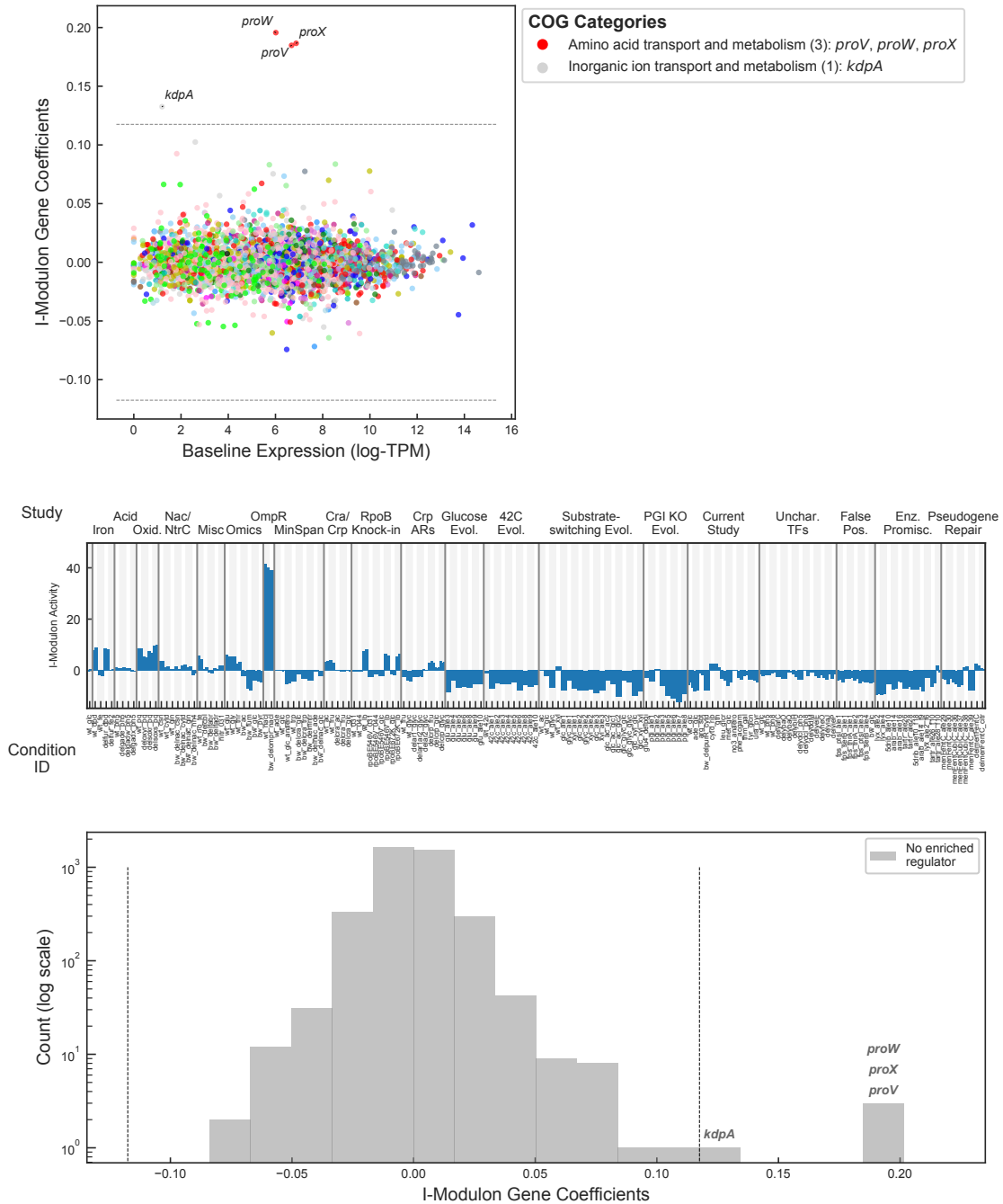
Biological Function: Nitric oxide response



Motif E-value: 5.30e-05
Operons with Upstream Motif: 75%

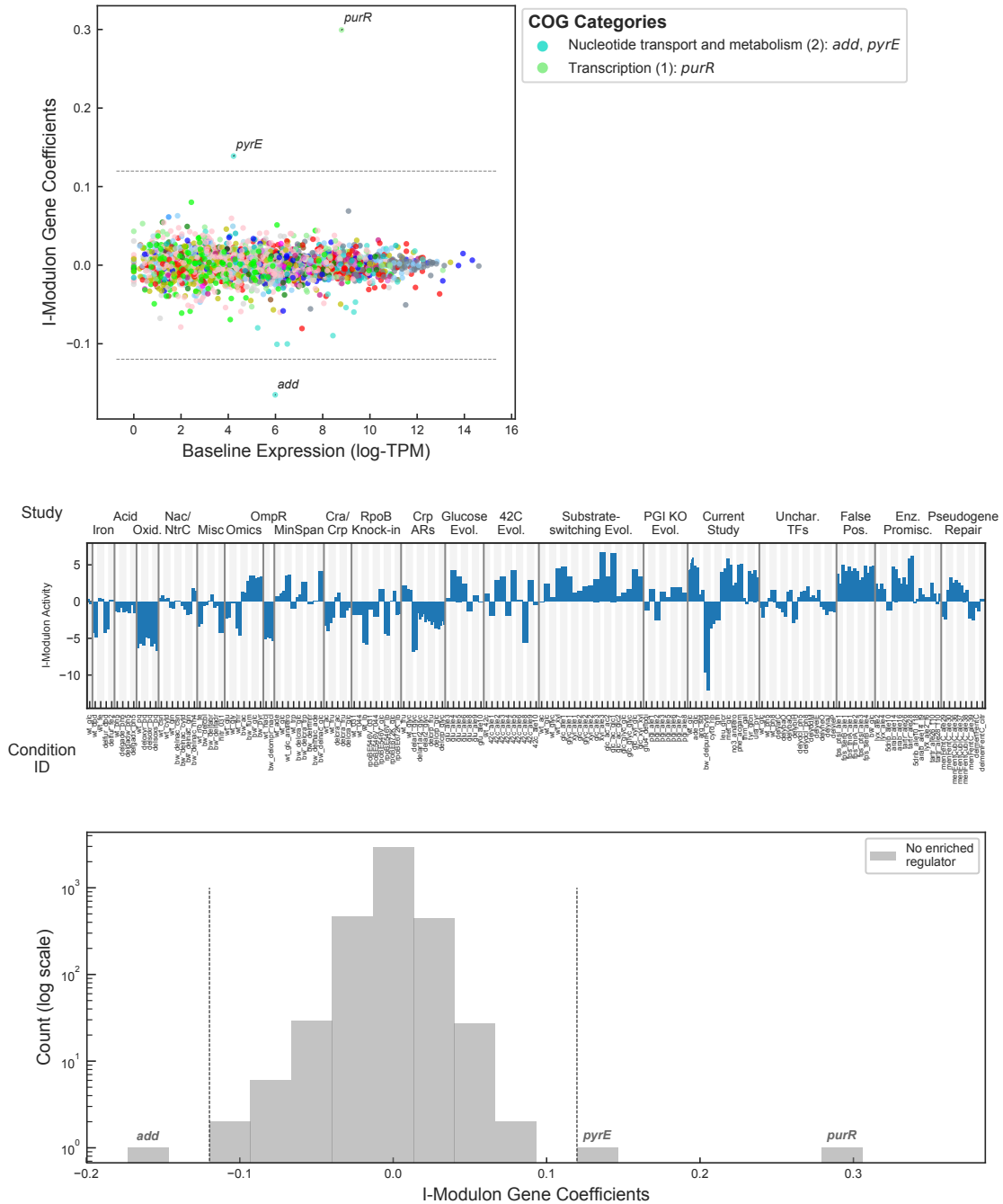
proVWX I-Modulon

Biological Function: Glycine betaine transport

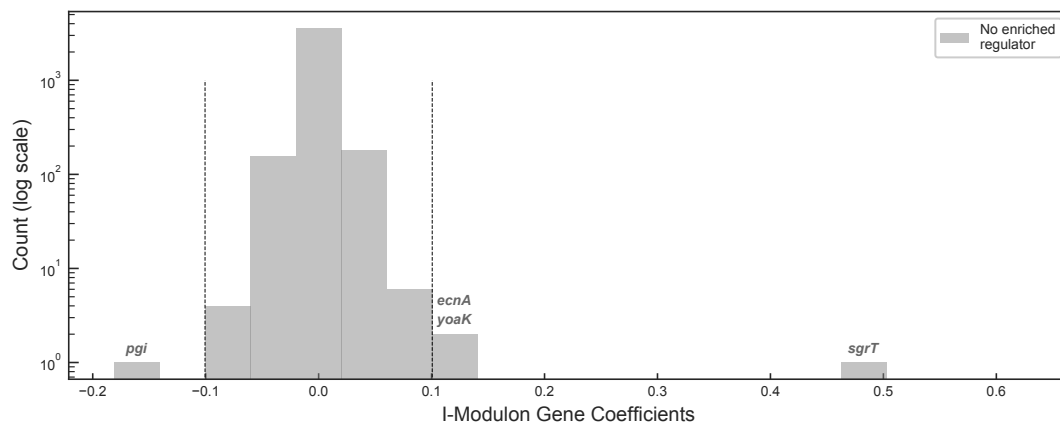
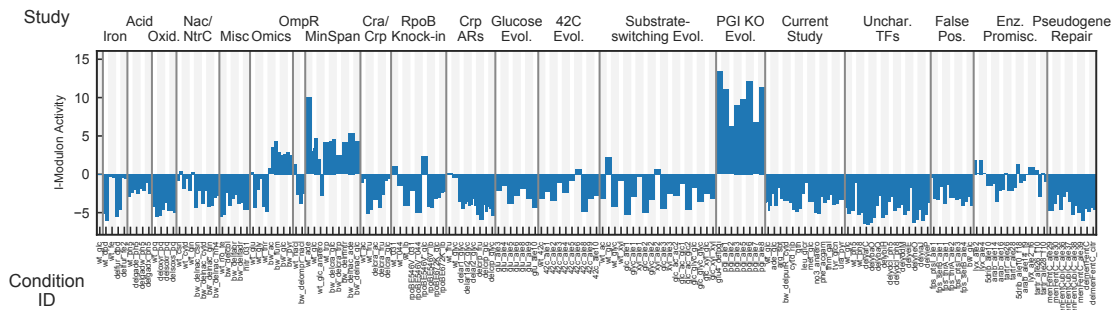
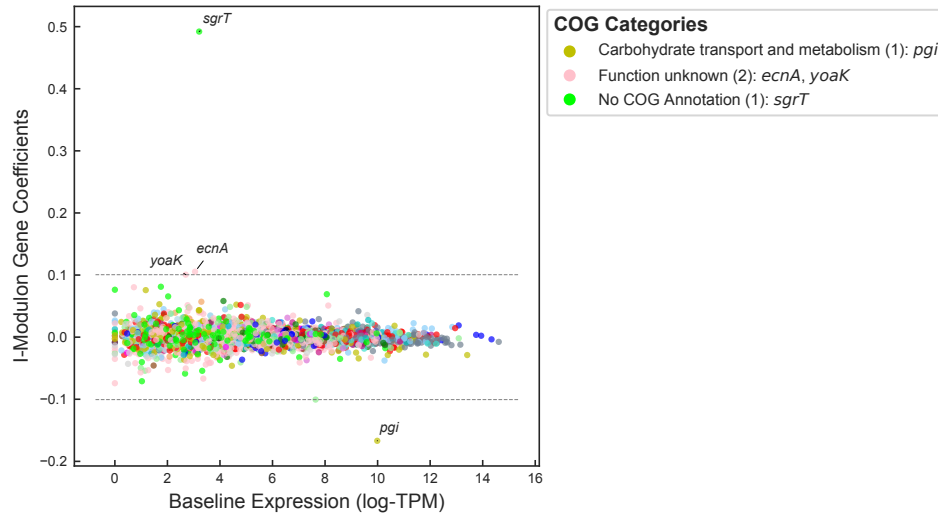


purR – KO I-Modulon

Biological Function: Accounts for purR knock-out

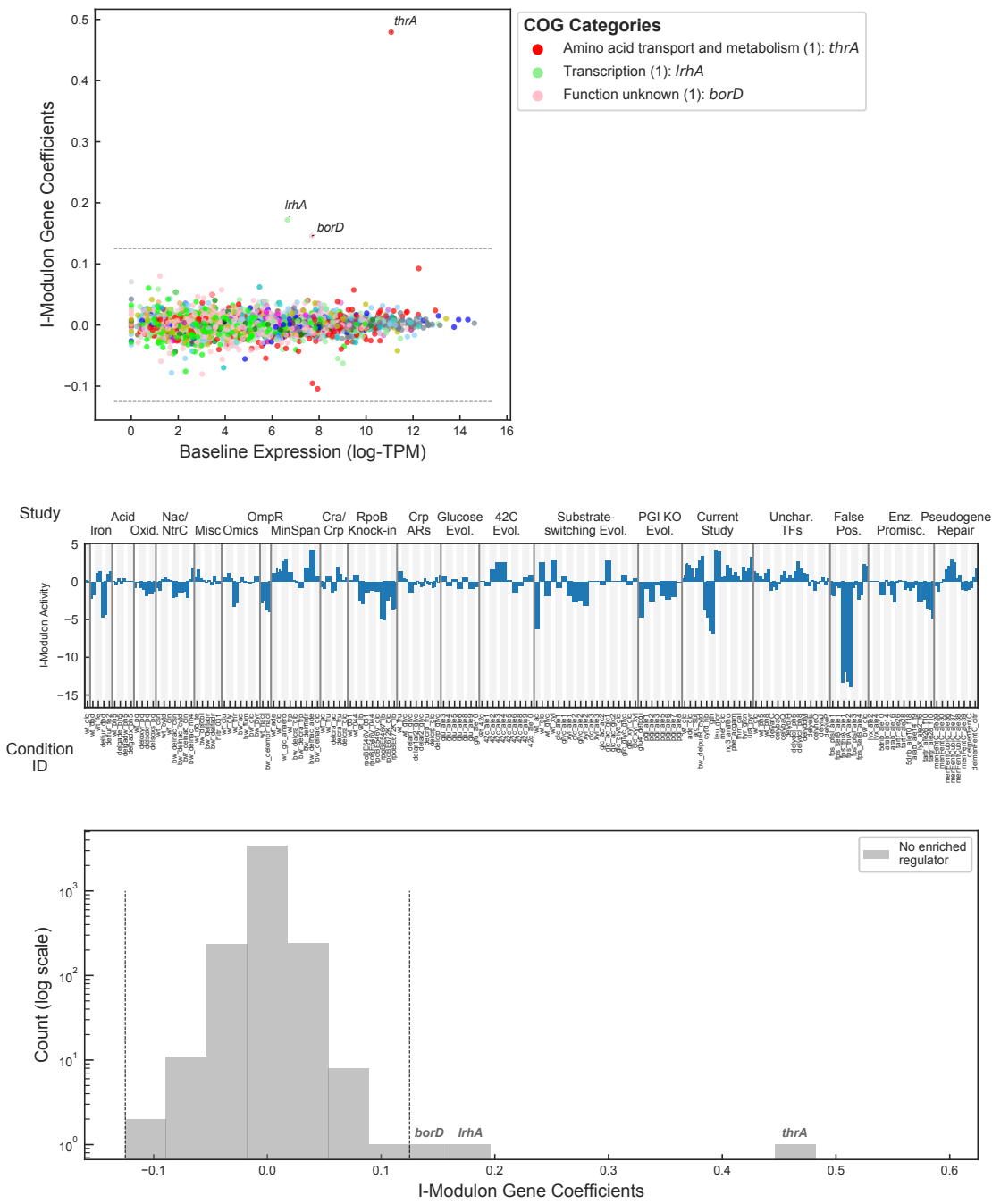


Biological Function: Contains single dominating gene: *sgrT*

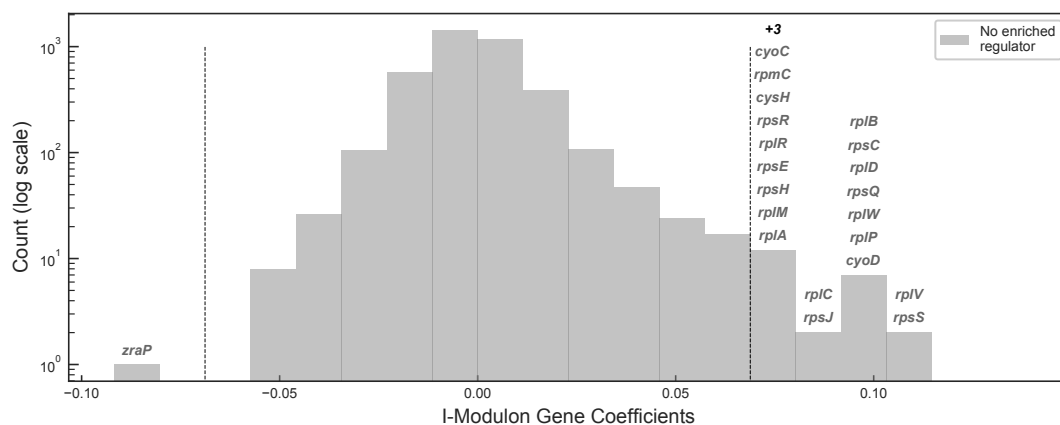
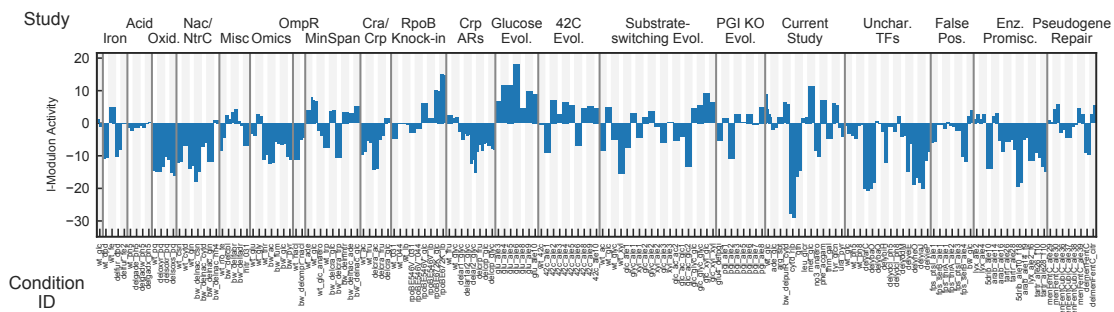
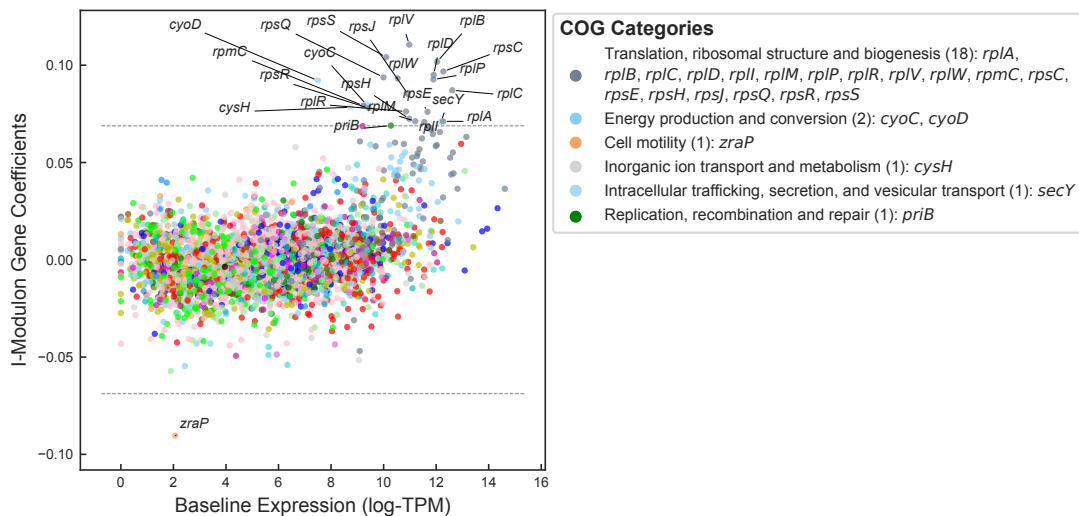


thrA – KO I-Modulon

Biological Function: Accounts for thrA knock-out

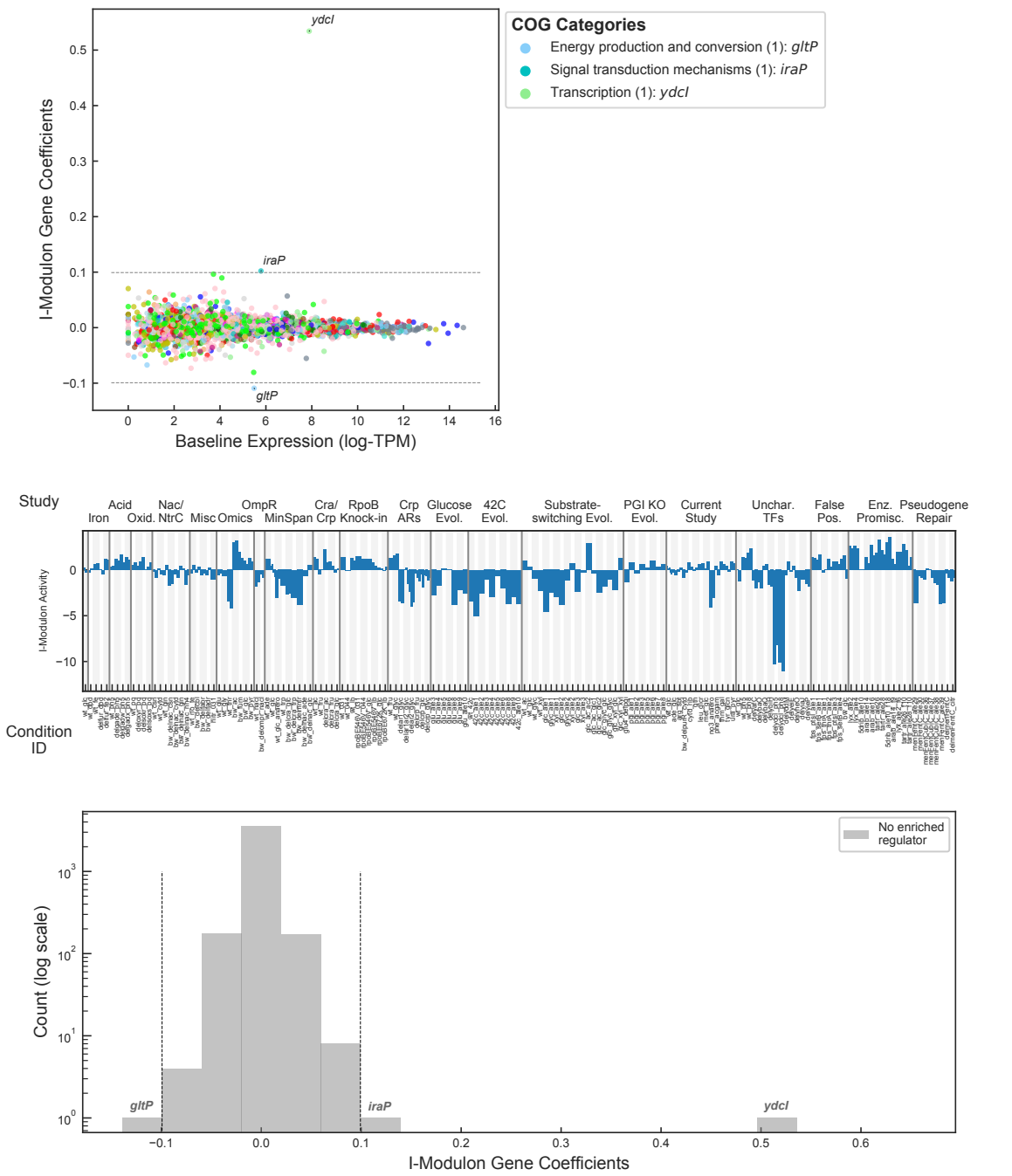


Biological Function: Enriched in translation machinery



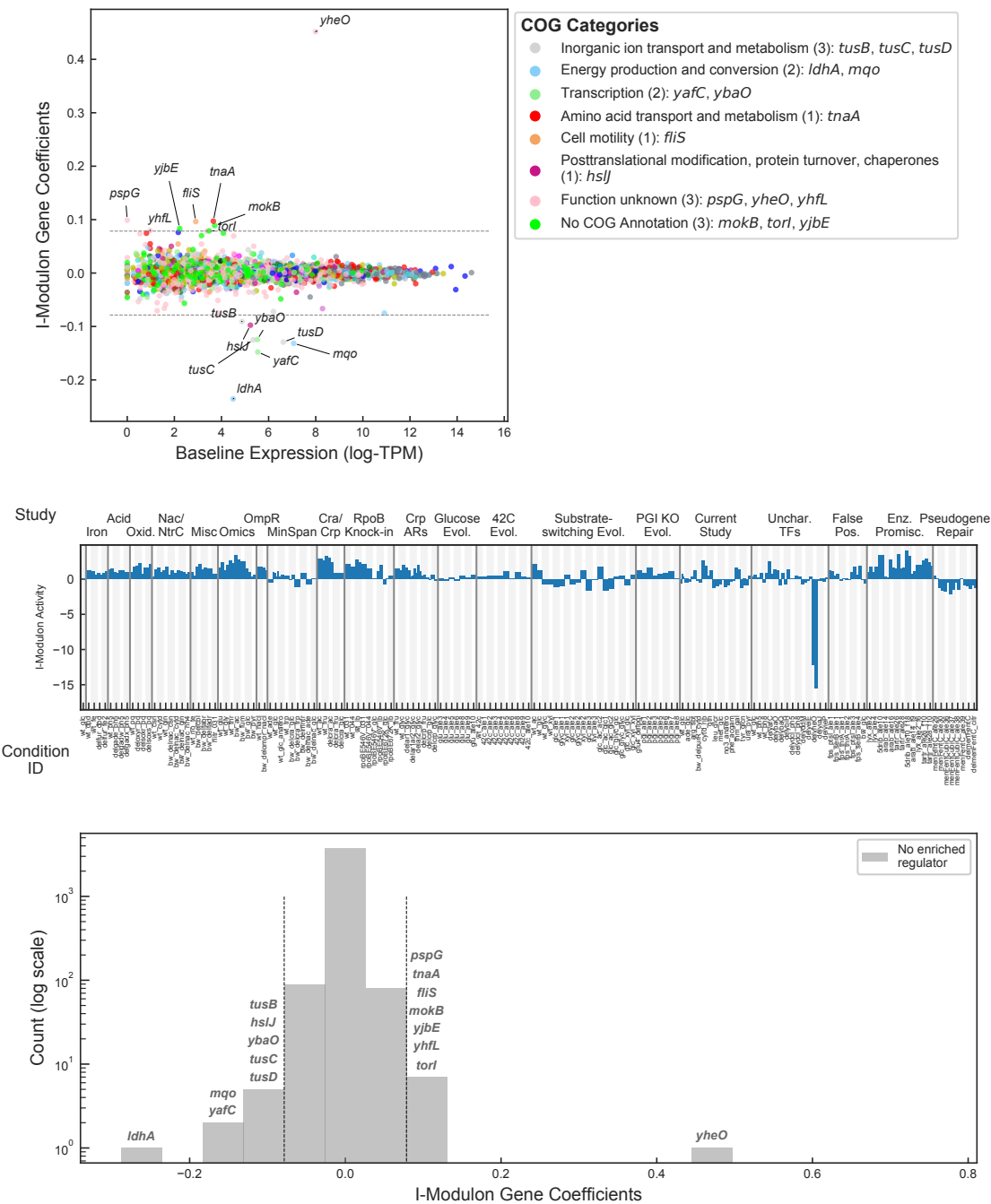
ydcl – KO I-Modulon

Biological Function: Accounts for ydcl knock-out



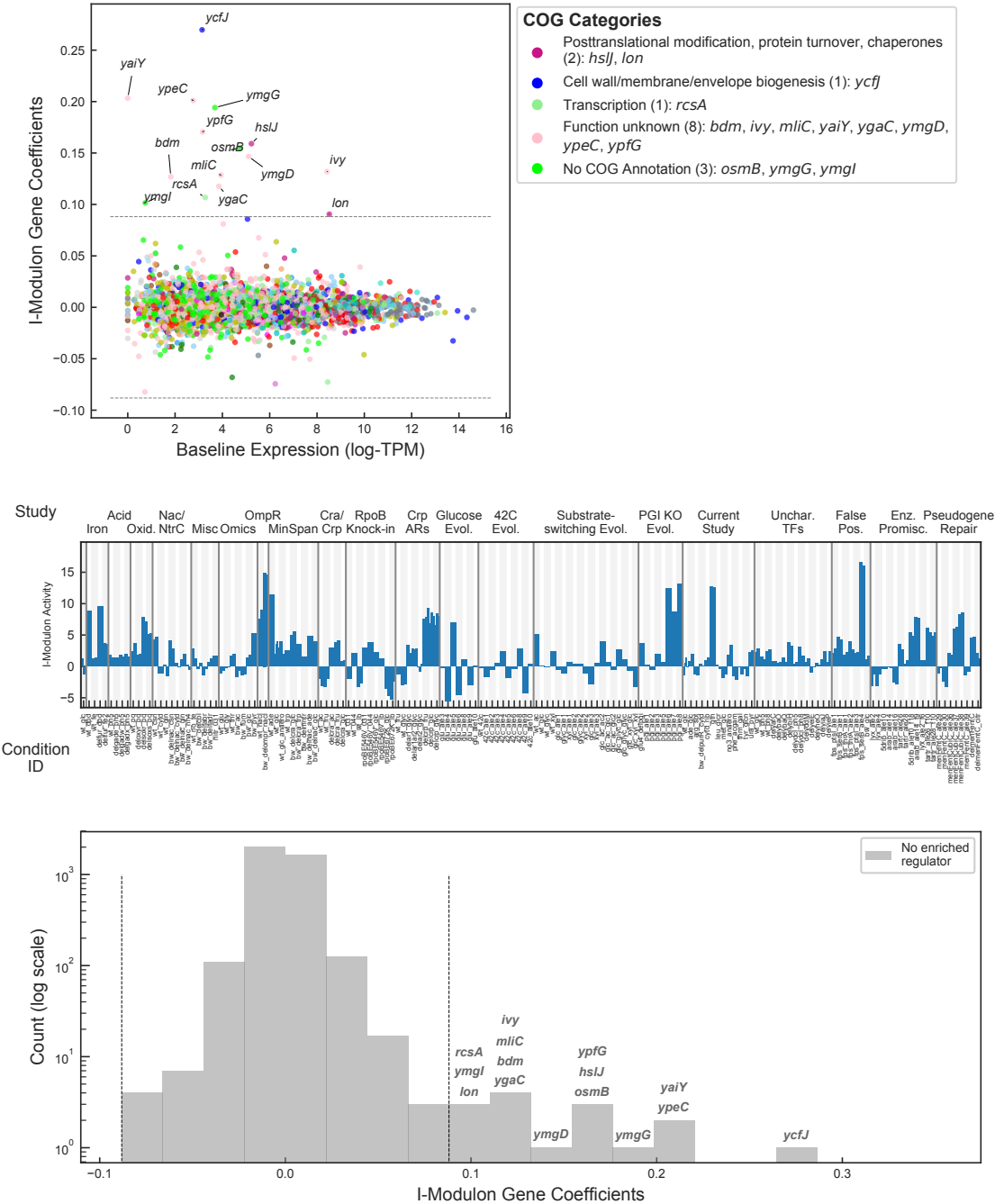
yheO – KO I-Modulon

Biological Function: Accounts for yheO knock-out



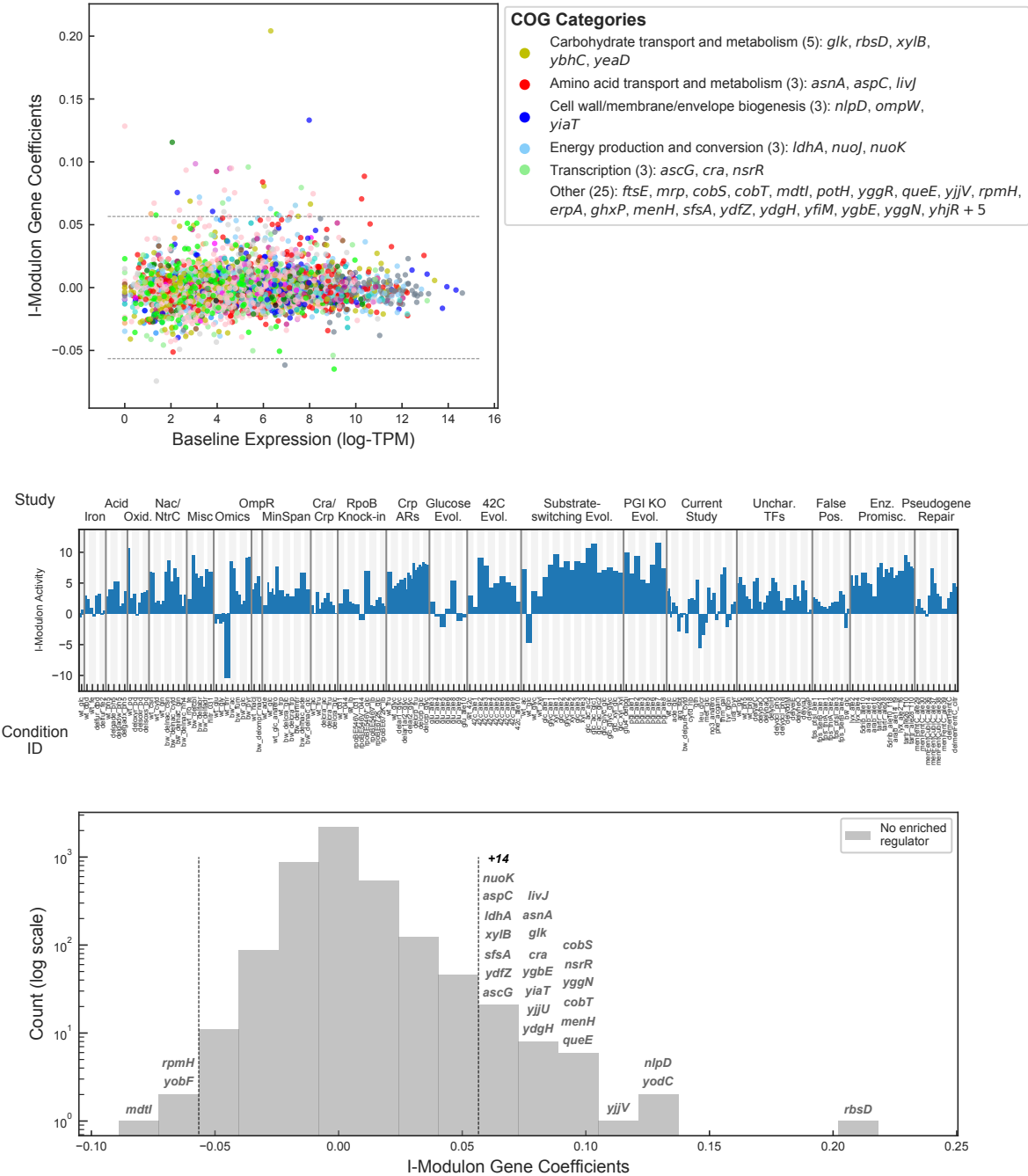
Uncharacterized I-Modulon #1

Biological Function: Unknown Function



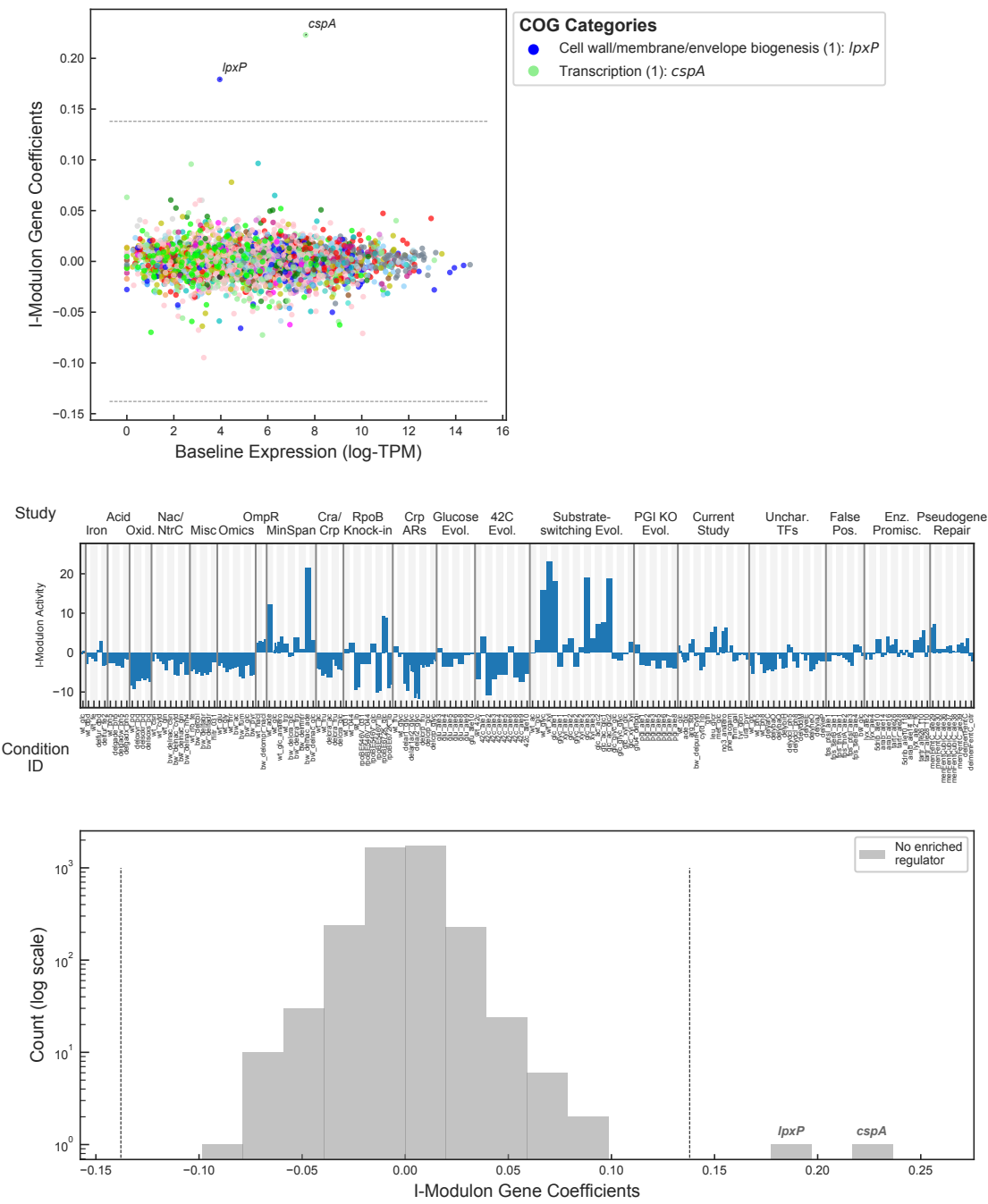
Uncharacterized I-Modulon #2

Biological Function: Unknown Function



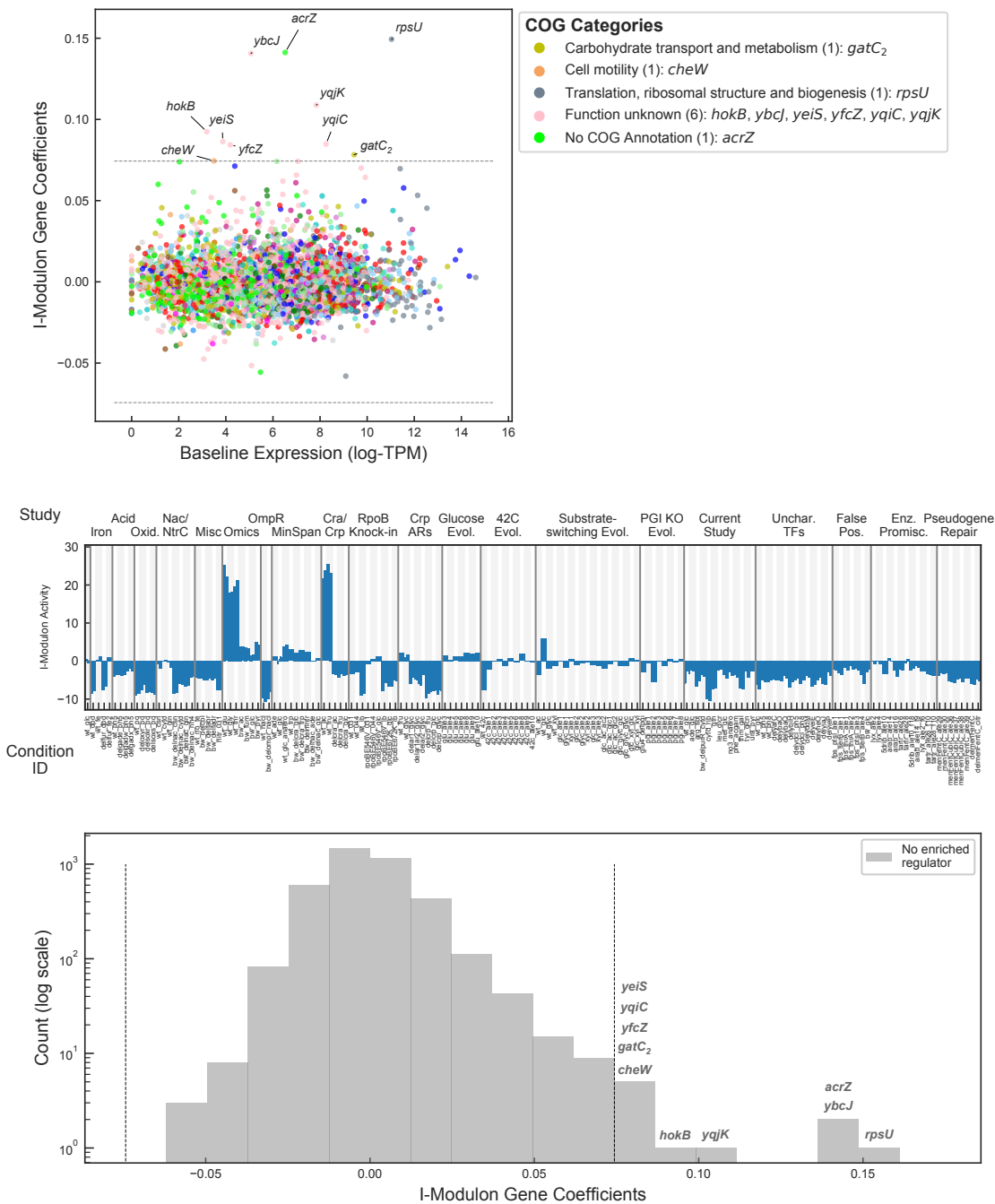
Uncharacterized I-Modulon #3

Biological Function: Unknown Function



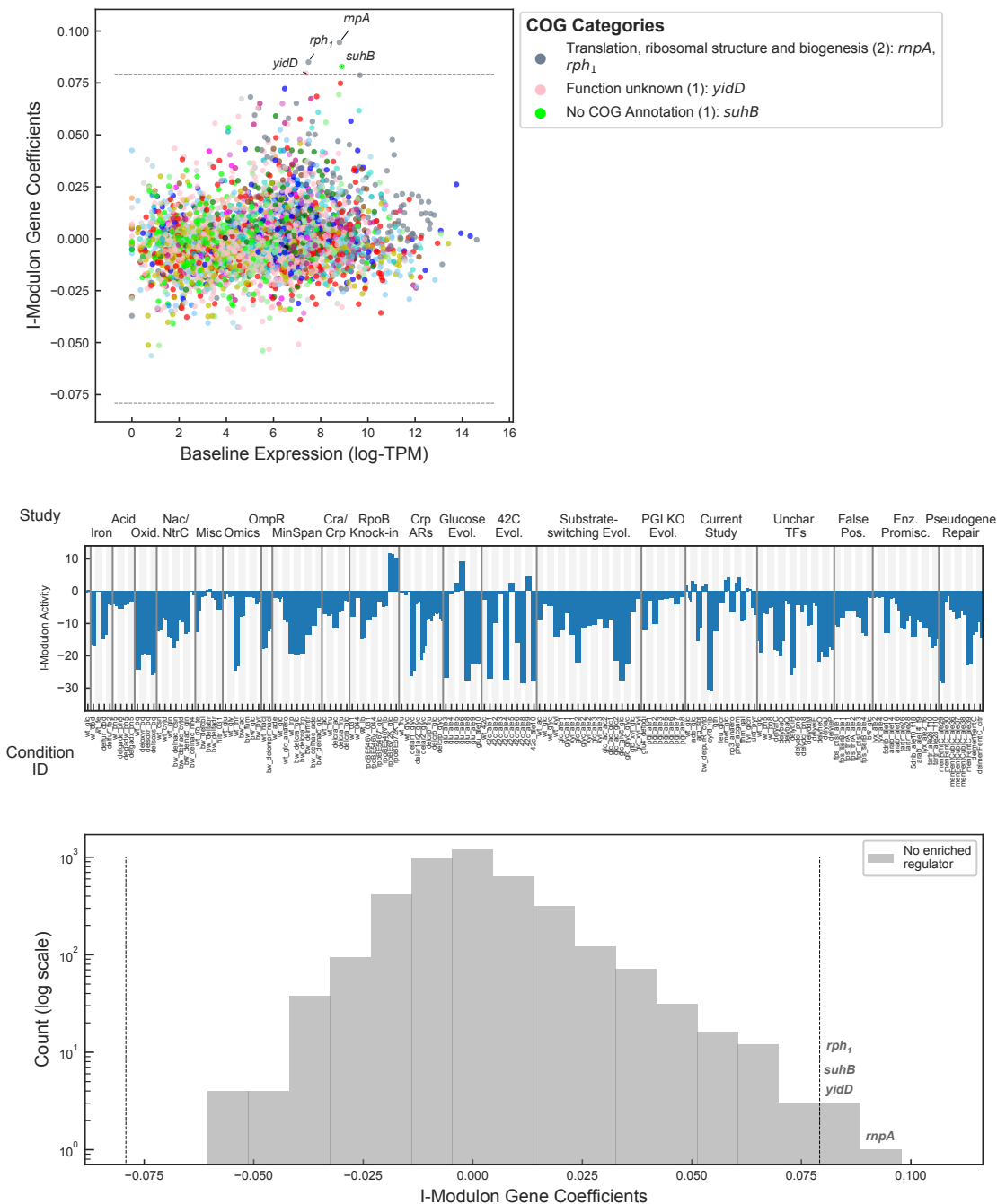
Uncharacterized I-Modulon #4

Biological Function: Unknown Function



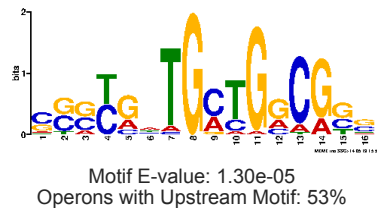
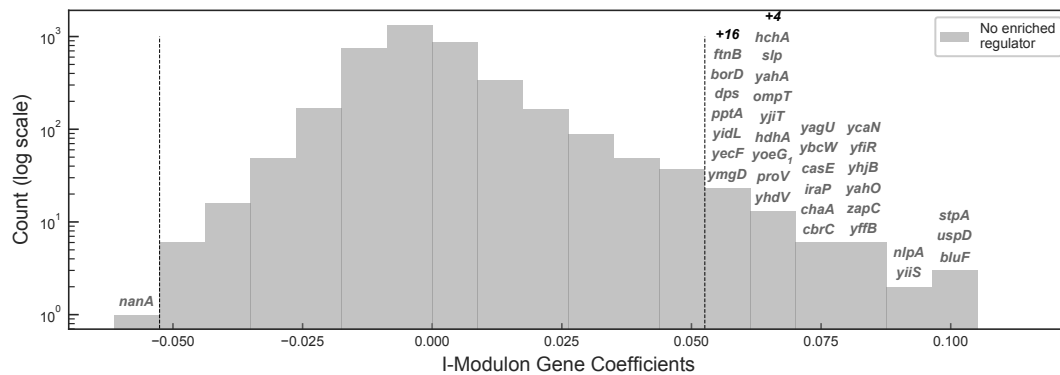
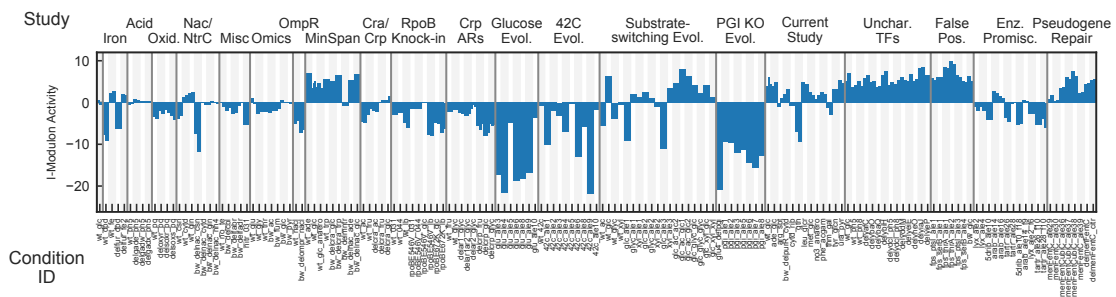
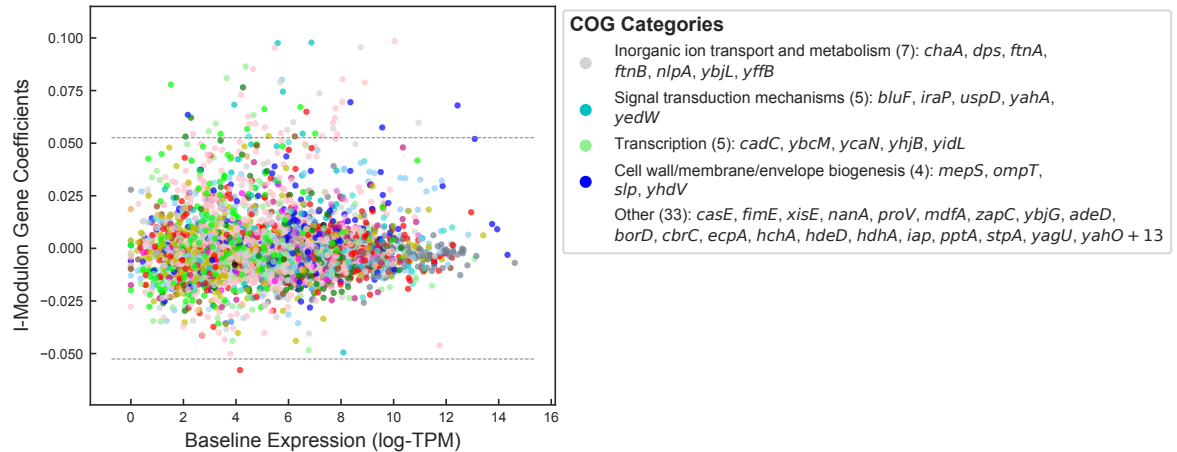
Uncharacterized I-Modulon #5

Biological Function: Unknown Function



Uncharacterized I-Modulon #6

Biological Function: Unknown Function



Appendix C

Matrix factorization recovers
consistent regulatory signals from
disparate datasets

C.1 Supplementary Figures

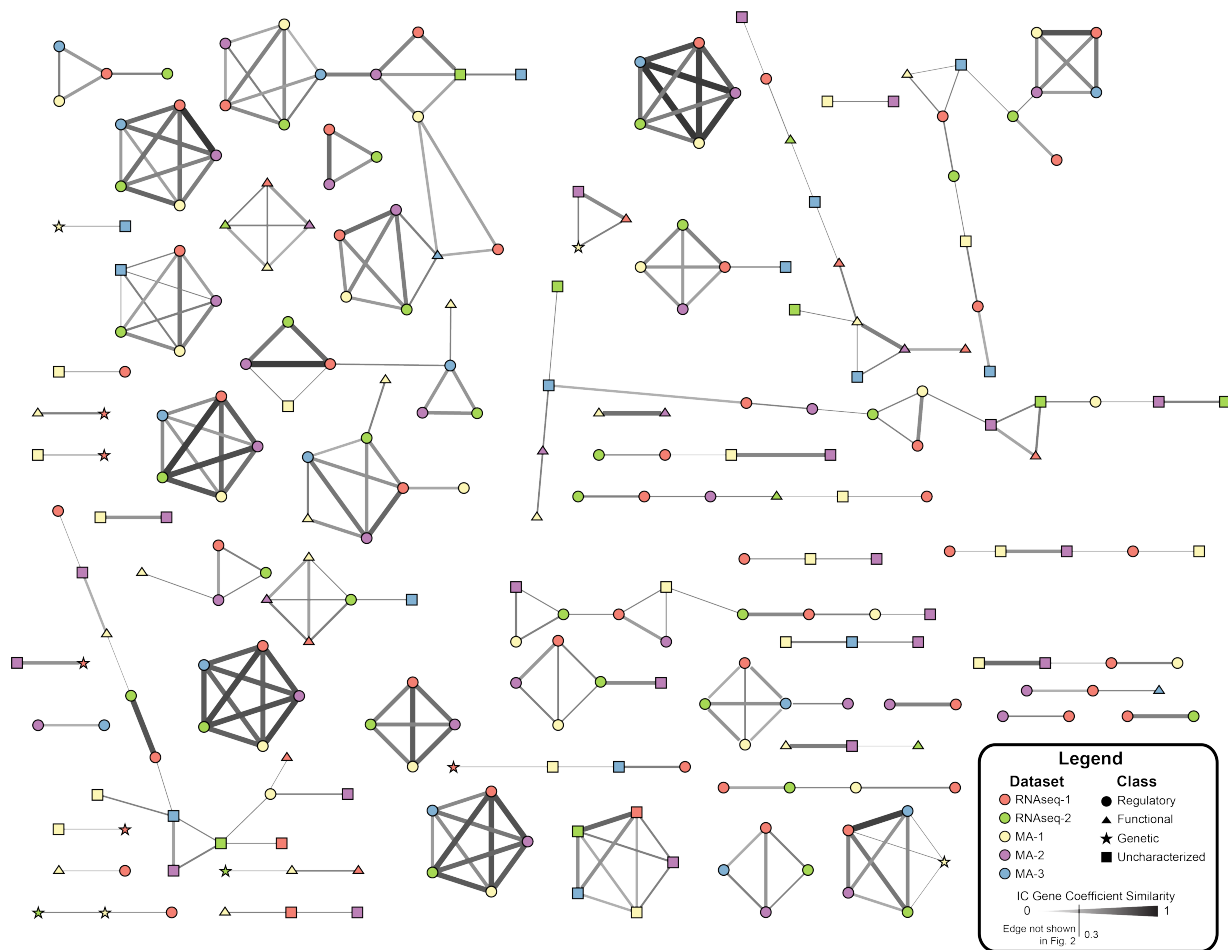


Figure C.1: Reciprocal best hit (RBH) graph containing all edges from the five datasets. Edges with an IC Gene Coefficient similarity score below 0.3 were pruned from Figure 3.3.

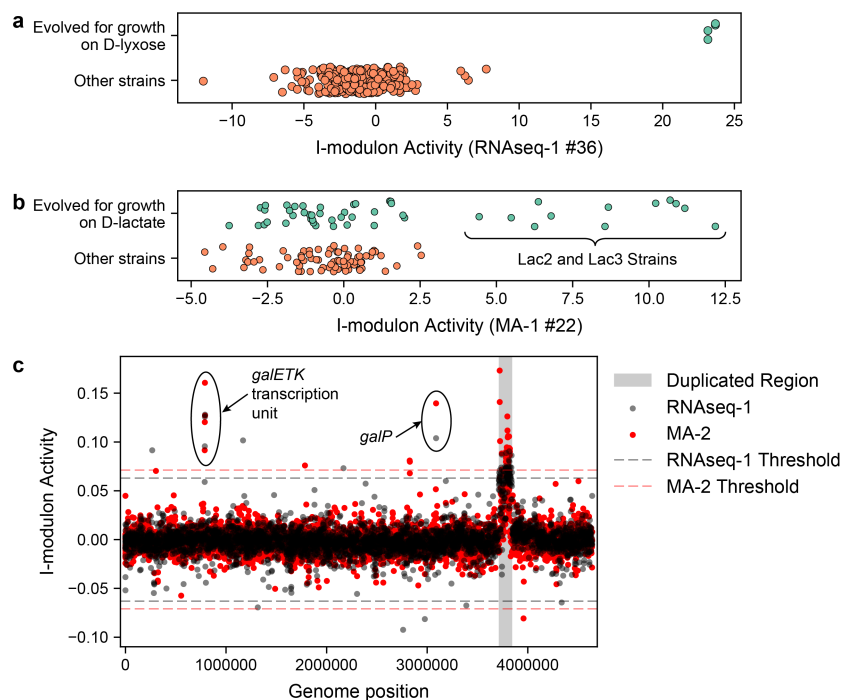


Figure C.2: Characteristics of the only Genomic i-modulon found in more than one dataset. (a) Activities of the i-modulon in the RNAseq-1 dataset separate *E. coli* strains evolved for growth on the non-native carbon source D-lyxose from the other strains in the compendium. (b) Activities of the uncharacterized i-modulon in MA-1 that was linked to the i-modulon described in panel (a). Seven strains were evolved in parallel for growth on D-lactate, but only two endpoint strains (named Lac2 and Lac3) exhibited high i-modulon activities. These strains were not resequenced, so the adaptive mutations cannot be confirmed. (c) Scatter plot showing the IC gene coefficients corresponding to the i-modulons described in panel (a) (in black) and panel (b) (in red). Thresholds determining i-modulon composition are indicated by dashed lines. The genomic duplication from the D-lyxose-evolved strains is highlighted in gray. Two transcription units outside of the duplicated region were captured in both i-modulons, the *galETK* transcription unit, and the *galP* gene. These genes are responsible for D-galactose catabolism.

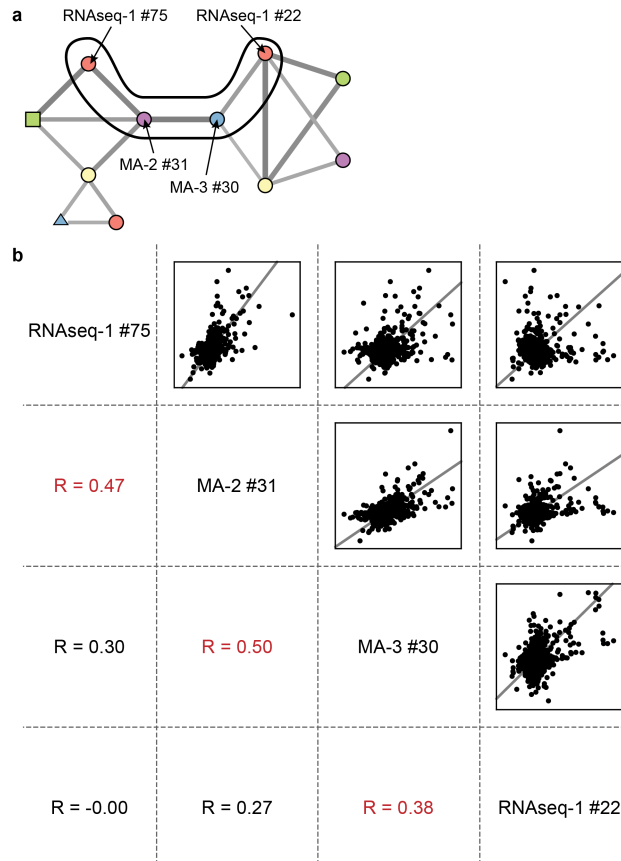


Figure C.3: Investigation of an irregular i-modulon cluster. (a) The irregular i-modulon cluster shows that two i-modulons from the RNaseq-1 dataset are indirectly connected. (b) Scatterplots of the IC gene weights for the i-modulons highlighted in (a). The pearson R correlation of the IC gene weights is shown below. The two RNaseq-1 i-modulons show no correlation, but still contain a few genes in common, indicating that the expression of these shared genes are controlled by two distinct underlying sources.

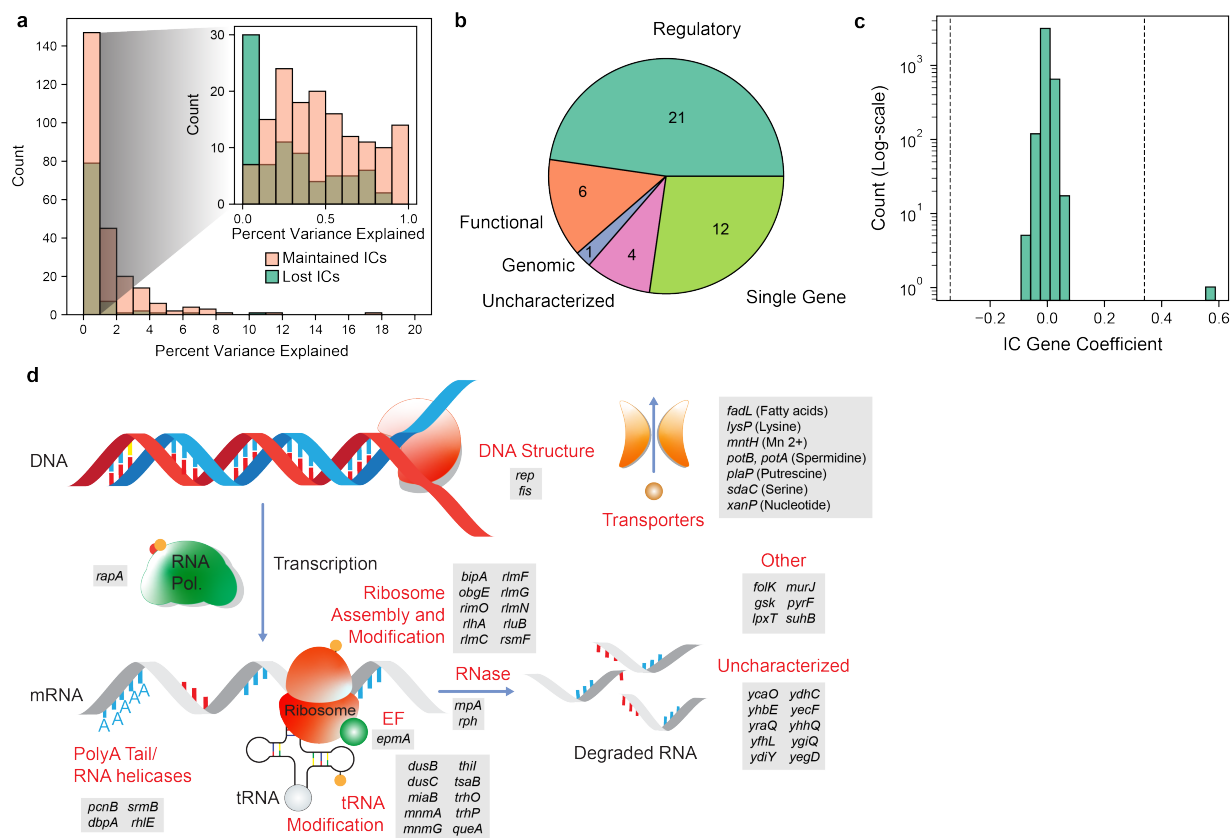


Figure C.5: ICA finds similar structures in the transcriptome and proteome. (a) Histogram of the percent of total expression explained by each independent component in the individual datasets. Components that are maintained (i.e. identify an RBH hit with the full compendium decomposition) are colored in pink, whereas components that are lost upon data integration are colored in blue. (b) Pie chart illustrating the classes of the new i-modulons extracted from the combined datasets. (c) Histogram of the IC gene coefficients of a Single Gene component. Dashed lines indicate the i-modulon threshold. (d) Schematic illustration of the various processes encoded by the genes in the Central Dogma i-modulon.

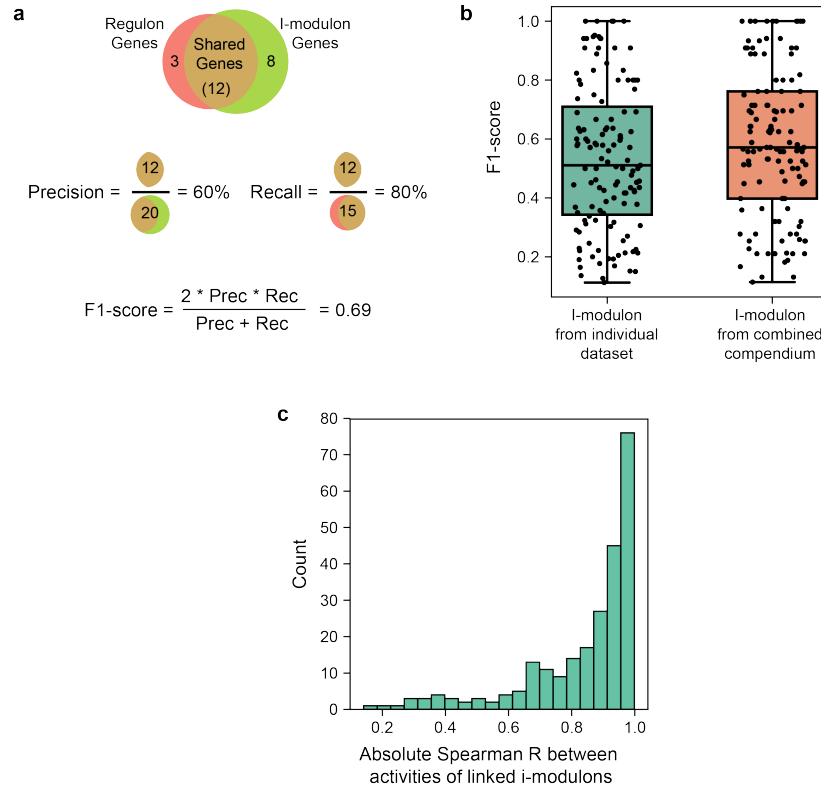


Figure C.6: Effects of data integration on i-modulon structure and activity. (a) Schematic illustration of precision, recall, and F1-score. (b) Boxplots of the F1-scores between i-modulons and their associated regulators for Regulatory i-modulons. Only Regulatory i-modulons in the individual dataset that found a reciprocal best hit in the full compendium are shown in the left boxplot, and the RBH of these i-modulons in the full dataset are shown in the right boxplot. (c) Histogram of absolute Spearman correlations between i-modulon activities in components that are RBHs in the full dataset compared to the individual datasets.

Appendix D

Decomposition of transcriptional
responses provides insights into
differential antibiotic susceptibility

D.1 Supplementary Figures

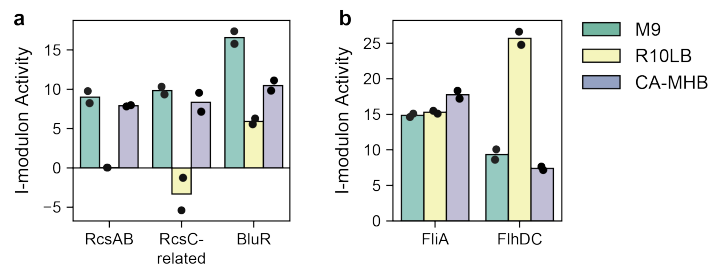


Figure D.1: Bar charts show the i-modulon activities for (a) Rcs-related i-modulons; (b) motility i-modulons. Individual measurements for biological replicates are plotted on top of bars. The RcsC-related i-modulon is discussed in Chapter 3.

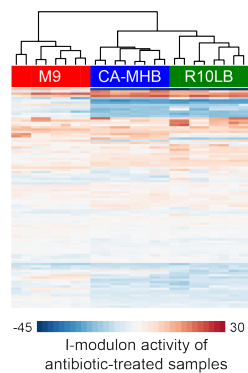


Figure D.2: Heatmap of i-modulon activities after antibiotic treatment.

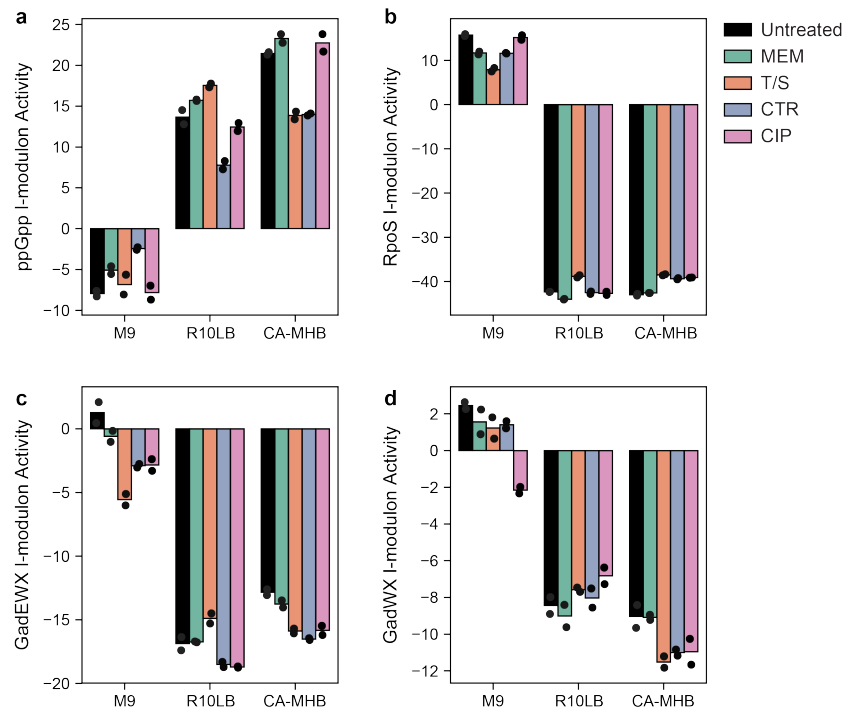


Figure D.3: Bar chart of fear-greed i-modulon activities, including the (a) ppGpp i-modulon; (b) RpoS i-modulon; (c) GadEWX i-modulon; (d) GadWX i-modulon.

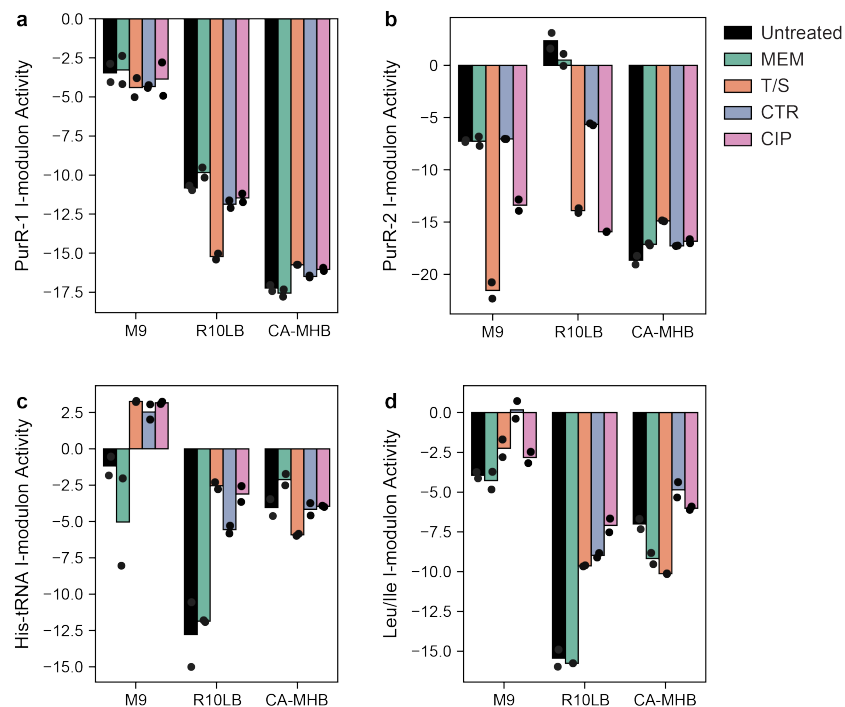


Figure D.4: Bar chart of nutrient-related i-modulon activities, including the (a) PurR-1 i-modulon; (b) PurR-2 i-modulon; (c) His-tRNA i-modulon; (d) Leu/Ile i-modulon.

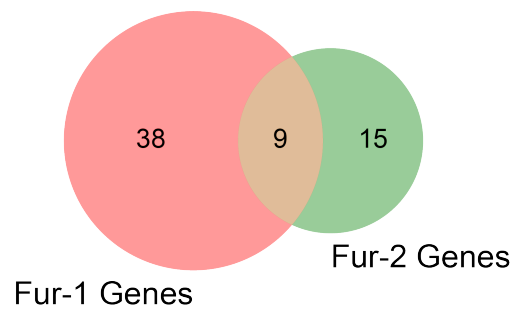


Figure D.5: Venn diagram of genes in the Fur i-modulons.