UNIVERSITY OF CALIFORNIA SAN DIEGO

Decomposition of *Synechococcus elongatus* transcriptomic data to reveal its regulatory modules through Independent Component Analysis

A thesis submitted in partial satisfaction of the

requirements for the degree

Master of Science

in

Bioengineering

by

Tahani Al Bulushi

Committee in charge:

Bernhard Ørn Palsson, Chair
Susan Stephens Golden
Anand Varun Sastry
Karsten B. Zengler

2021

The thesis of Tahani Al Bulushi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To all the supportive women in my family,

my grandma, my mother, and my sister,

without whom I would not be where I am today.

This work is for you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Bernhard Palsson for his continuing support and guidance throughout the progression of my thesis. His scientific experience and valuable advice motivated me to build confidence in my research. Regardless of his busy schedule, he accommodated my inquiries and provided me with all the necessary resources so I could deliver my work without any hurdles. He continues to motivate me and I am grateful to have him as my advisor.

I also want to thank the great mentors I had while at SBRG. First, I would like to thank Anand Sastry for teaching me every skill I have learned in the lab, from building a statistical understanding of biology and implementing different computational tools. His scientific insight and deep understanding of systems biology have not only enabled me to grasp these concepts but also motivated me to learn them in a short amount of time. He is truly a great role model, project leader, and a kind human being. Next is Kevin Rychel who was the main inspiration to learn the power of big data and programming. His passion for computational learning and teaching skills were exemplary and motivated me to join SBRG.

I am also grateful for the valuable discussions I have had with Dr. Susan Golden, Cigdem Sancar, and Marie Adomako. They have helped me understand cyanobacteria to a great level of detail. I have also been fortunate to have great peers in the lab; especially Saugat, Sidd, Reo, and Annie. I am thankful to have gained your company through multiple weekly meetings even though I have not had the pleasure of meeting most of you in person due to COVID-19.

I am especially thankful for the love, inspiration, and support from my mother and sister. I would not be standing where I am today without them and for that, I am forever grateful.

Chapter 1, 2, 3, and 4 are currently being prepared for submission for publication of the material. **Tahani Al Bulushi**, Anand V Sastry, Kevin Rychel, Saugat Poudel, Reo Yoo, Siddharth Chauhan, Yuan Yuan, Cigdem Sancar, Richard Szubin, Bernhard Ø. Palsson, Susan Golden. (2021). “Machine learning reveals the transcriptional regulatory network and circadian dynamics of the cyanobacteria *Synechococcus elongatus* PCC 7942”. The thesis author is the primary author.

Chapter 3, in part, is material submitted for publication. Anand Sastry, Saugat Poudel, Kevin Rychel, Reo Yoo, Cameron Lamoureux, Siddharth Chauhan, Zachary B. Haiman, **Tahani Al Bulushi**, Yara Saif, and Bernard Ø. Palsson. (2021) Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. BioRxiv. DOI: https://doi.org/10.1101/2021.07.01.450581. The thesis author is the co-author.

ABSTRACT OF THE THESIS

Decomposition of *Synechococcus elongatus* transcriptomics data to reveal its regulatory modules through Independent Component Analysis

by

Tahani Al Bulushi

Master of Science in Bioengineering

University of California San Diego, 2021

Professor Bernhard Ø. Palsson, Chair

Synechococcus elongatus is a tractable model cyanobacterium for circadian studies and a platform for bioproduction. The organism's adaptation response to conditional changes in aquatic environments is orchestrated through the transcriptional regulatory network (TRN). Despite the previous characterization of constituent parts of the *S. elongatus* TRN, a system-level characterization and analysis of the interactions between major transactional regulators have yet to be established. Here, we demonstrate the utility of unsupervised machine learning to compartmentalize and describe the characteristics of the different regulatory modules of the model strain *S. elongatus PCC 7942,* enabling a complete reconstruction of its TRN in response to environmental stresses and changes in intracellular states. Through the application of Independent component Analysis (ICA) to a collection of 317 transcriptomic samples, we obtained 51 independently modulated gene sets called “iModulons", each of which explained a portion of the variance in the organism's transcriptional response. iModulons serve as a knowledge tool to elucidate the transcriptional function and activation dynamics of previously undefined regulons while also describing the interaction between transcription factors in the TRN. Our data-driven analysis also provides, for the first time, a complete TRN reconstruction for *S. elongatus* with valuable functional context to expand the annotation of many hypothetical genes captured in our iModulon structure. This transcriptome-wide analysis of *S. elongatus* TRN informs future research on areas of possible genetic perturbations to manipulate its transcriptional regulation and optimize the engineering of this organism. A knowledge-driven database of all published high-quality RNA-seq data for *S. elongatus* to date is now available in iModulonDB.org.

Chapter 1
Introduction

1.1 Background

Microbes have the ability to adapt to diverse environmental changes by coordinating their gene expression state through a complex system called the transcriptional regulatory network (TRN). In bacteria, a TRN constitutes a wide range of transcriptional regulatory units, including proteins like transcription factors (TF) (Ishihama, 2000) and sigma factors, and RNA fragments like riboswitches (Nudler & Mironov, 2004), small RNAs (Gottesman & Storz, 2011), and transcriptional attenuators (Yanofsky, 1988). Biological methods such as chromatin immunoprecipitation (ChIP)(Rhee & Pugh, 2012) are capable of identifying the binding sites of transcription and sigma factors for genome-wide discovery. The TRN behaves similarly to a signal processing unit. It participates in signal transduction pathways by sensing environmental and intracellular signals, processing signal information, then applying appropriate adjustments to gene expression to optimize bacterial growth and survival (**Figure 1.1**). The most well-studied TRN in bacteria is that of *Escherichia coli*, an intestinal bacteria that modulates its gene expression depending on levels of nutrient and oxygenation present in the host environment (Jones et al., 2007). RegulonDB documents over 7,000 interactions between transcription regulators and *E coli's* genetic material (Santos-Zavaleta et al., 2018). The development of high throughput RNA sequencing (RNA-seq) technologies has facilitated the collection of high-quality microbial gene expression datasets, or transcriptomes, that can be analyzed to reconstruct microbial TRN and elucidate their functional involvement. The reconstruction of or reverse engineering of the TRN components informs the means to how an organism responds to diverse environmental stresses and unfamiliar conditions (Buescher et al., 2012). Once the TRN is reconstructed, it can be characterized to enable data-driven predictions that elucidate the dynamics of the different adaptation responses as a result of environmental and genetic alterations. To enable TRN reconstruction, a substantial amount of experimental transcriptomic data is needed to exploit all the binding sites for each DNA-binding regulatory and study their condition-dependent (Sastry et al., 2019).

The TRN provides an important window to understanding many issues surrounding health care and biotechnology. Transcription factors have been shown to control antibiotic resistance in *Salmonella spp.* (Bailey et al., 2010), *Salmonella Typhimurium* (Abouzeed et al., 2008; Alekshun & Levy, 1997), and *Neisseria gonorrhoeae* (Zalucki et al., 2012), virulence in pathogenic bacteria like *Mycobacterium tuberculosis* (Raghavan et al., 2008), and the development of engineered strains for bioproduction processes (Choe et al., 2019; Mohamed et al., 2017). Therefore, a comprehensive understanding of TRN structures would enable the prediction and elucidation of transcriptomic changes in response to multiple environmental cues and intracellular states.

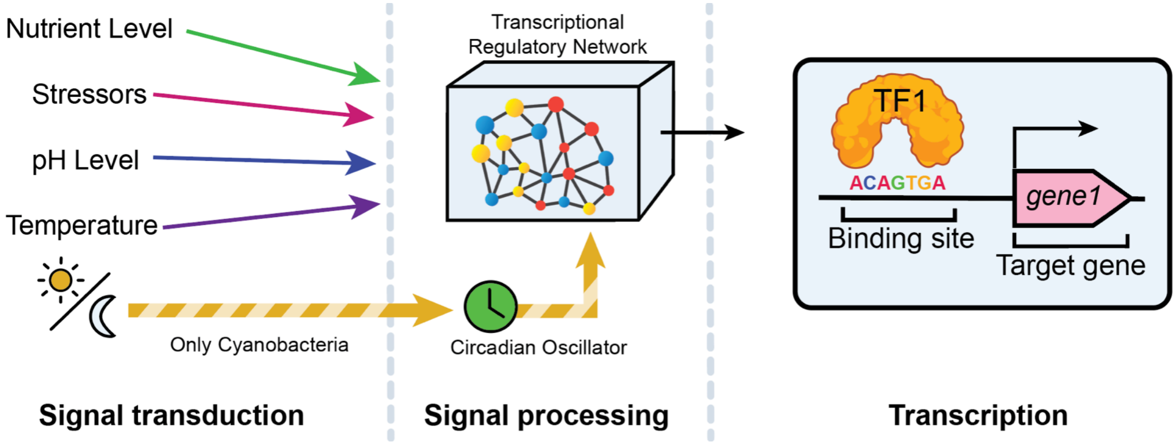


Figure 1.1: A visual representation of the TRN function. The TRN receives extracellular and intracellular signals to influence transcriptome dynamics through gene expression as an important method for cellular adaptation. Shown by the yellow arrow is the only exception in TRN regulation observed in most cyanobacteria; where the activity of the endogenous clock, or circadian oscillator, is directly influenced by light and dark cycles.

1.2 The TRN of *Synechococcus elongatus* PCC 7942

Similar to *E. coli*, cyanobacteria also incorporate a responsive TRN to fluctuating environmental conditions such as osmolarity (Paithoonrangsarid et al., 2004; Shoumskaya et al., 2005), salinity (Marin et al., 2003), and temperature (Suzuki et al., 2001) variations. Compared to many well-studied heterotrophic bacteria, such as *E. coli* or *Bacillus subtilis*, cyanobacteria remain relatively understudied. These microorganisms are a photoautotrophic prokaryotic group capable of performing oxygenic photosynthesis. Because of their high dependency on photosynthetic precursors such as light and carbon dioxide (CO_2), their TRN significantly differs from previously characterized heterotrophs. The majority of cyanobacterial transcription regulators are influenced by its light-sensitive circadian clock mechanism, which governs fundamental cellular processes such as central metabolism, photosynthesis, and growth. Therefore, it is expected that most TF in cyanobacteria is regulated either directly or indirectly by the circadian clock (**Figure 1.1**) (Nair et al., 2001; Tu et al., 2004).

A model strain for prokaryotic circadian studies in cyanobacteria is *Synechococcus elongatus* PCC 7942 (henceforth referred to as *S. elongatus*) because of its tractable and small genome size (approximately 2.7 Mb), efficient homologous recombination, and its viability of saturation mutant screens (Andersson et al., 2000). Its main circadian oscillatory protein (KaiC) has been shown to physiologically change its composition in response to light-dark (LD) cycles every 24 hours, regulate gene expression through clock output proteins, and influence metabolic shifts such as that between glycolysis and gluconeogenesis (Diamond et al., 2015). Despite this, *S. elongatus* lacks a comprehensive TRN structure that encompasses the majority of its cellular processes (Minezaki et al., 2005). For example, the BioCyc Database Collection reports nine TFs for *S elongatus*, which is not updated with recent literature findings in terms of observed or proposed regulatory units.

Since *S. elongatus* base their metabolism on photosynthesis, almost all cellular processes are affected by the presence or absence of light, which presents a unique engineering challenge for industrial applications. *S. elongatus* have gained increasing attention from the field of biotechnology due to their efficiency as photoautotrophs in converting CO_2 to useful industrial biochemicals, tolerance to high temperatures, and ability to harness solar light as their main source of energy, reducing production costs. Because of these desirable characteristics, *S. elongatus* has been engineered to produce ethanol (Deng & Coleman, 1999), isobutyraldehyde (Atsumi et al., 2009), alkanes (Schirmer et al., 2010), hydrogen (Kruse & Hankamer, 2010), and free fatty acids (Ruffing & Jones, 2012) as biodiesel precursors in large-scale applications. Other studies also show the important role of *S. elongatus* in limiting biological contaminants by using phosphite in outdoor cultivation systems (National Research Council et al., 2013). However, since most of these desired features are affected by LD cycles, it is important to understand how the TRN coordinates gene expression to optimize metabolic goals.

1.3 Machine Learning and Transcriptomic Datasets

The Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) is a public repository for high-throughput functional genomic datasets, maintained by the National Center for Biotechnology Information (NCBI) (Barrett et al., 2012). RNA-seq data availability for *S. elongatus* has increased within recent years, (**Figure 1.2**) motivating us to utilize the public data to investigate the TRN structure for *S. elongatus*. We processed the downloaded data using our RNA-seq processing pipeline followed by our quality control (QC) pipeline to generate 317 high-quality RNA-seq expression profiles for *S. elongatus* (see chapter 3 for Methods).

In the era of big biological data, the demand for robust computational tools using statistical and machine learning techniques is on the rise. The complexity of biological data can be made comprehensible for downstream analysis through computational tools to derive new biological knowledge. Unlike supervised machine learning, which requires an expected outcome, unsupervised machine learning distinguishes hidden structures from an unlabeled dataset. One of the major categories of unsupervised learning is matrix decomposition, which applies the blind signal separation approach to decompose a high-dimensional matrix into two lower-dimensional feature-describing matrices of the separated signals. Common matrix decomposition algorithms include principal component analysis, independent component analysis, non-negative matrix factorization, and singular value decomposition. This thesis applies independent component analysis on the gene expression dataset to produce two matrices as described in the next section.

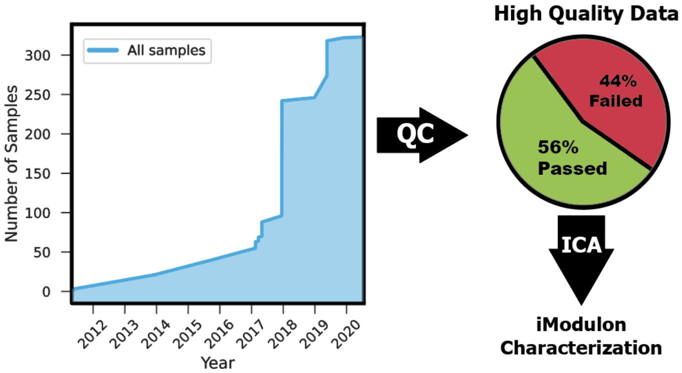


Figure 1.2: Publically available RNA-seq data for *S. elongatus* and quality control. (a) A line chart representing the accumulation of publicly available RNA-seq dataset for *S. elongatus* in NCBI (total 323) as of August 2020 **(b)** The dataset was processed using our quality control (QC) pipeline that applies five statistical criteria (**Figure A.1a**, see chapter 3 for more details). 56% (or 177 samples) of the dataset passed the QC pipeline and 44% (or 140 samples) of the dataset was discarded.

1.4 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a decomposition algorithm. It is used to decompose the original mixed signal into their constituent individual elements and determine their relative strengths. A common analogy to describe ICA is the “cocktail party problem” which refers to the ability to follow a single auditory signal from one person while filtering out other sources of noise from the party room (Brown et al., 2001). If ICA was applied to the recording of linear mixed auditory signals produced by a set of sources, the algorithm exploits a statistical discriminant to differentiate these sources and separate them in a “blind” manner. The mathematical formula for this decomposition is represented by the equation **X = MA,** where multiplying both **A** and **M** matrices will result in the reconstruction of the **X** expression profile matrix. ICA is built on the assumption that individual source signals are statistically independent of each other, meaning that every signal shows no effect on other signals and that the values in each source signal have a non-Gaussian distribution. With this, ICA maximizes statistical independence through the measure of non-Gaussianity using kurtosis and negentropy, which measures the distance from normality (Hyvärinen et al., 2004). With ICA gaining a lot of attention, the algorithm was heavily used in the disciplines of signal processing and neural computation which led to the development of FastICA.

Recently, ICA has been applied to a high-quality RNA sequencing (RNA-seq) gene expression compendium (Poudel et al., 2020; Rychel et al., 2020; Sastry et al., 2019) to isolate independently regulated transcriptional modules (named "iModulons") for E. coli. This method was then applied to expand the current understanding of the *Bacillus subtilis* and *Staphylococcus aureus* TRNs. As described in the previous section, ICA decomposes a compendium of gene expression data into two matrices: the iModulon matrix (**M**) describing the relationship between gene affected by an underlying biological signal, and the Activity matrix (**A**) containing condition-specific activity levels for each iModulon (**Figure 1.3**). Each iModulon comprises a set of genes whose expression varies collectively, but independently of other genes not present in the given iModulon. Thus, iModulons represent functionally-related coexpressed gene sets across multiple conditions. The activities encoded in the **A** matrix elucidate the changes in expression levels for the collective genes comprising a single iModulon under different conditions.

Prior studies have applied similar techniques to cluster significantly expressed genes in clusters or networks to map out the transcriptional topology in cyanobacteria (Singh et al., 2010; Yang et al., 2015). For example, Context-Likelihood of Relatedness (CLR) was used to create a network of co-expressed genes organized in graph neural networks (McClure et al., 2016).

Methods like CLR and ICA are extremely beneficial in characterizing understudied photoautotrophs because of their lack of operon-based genetic organization.

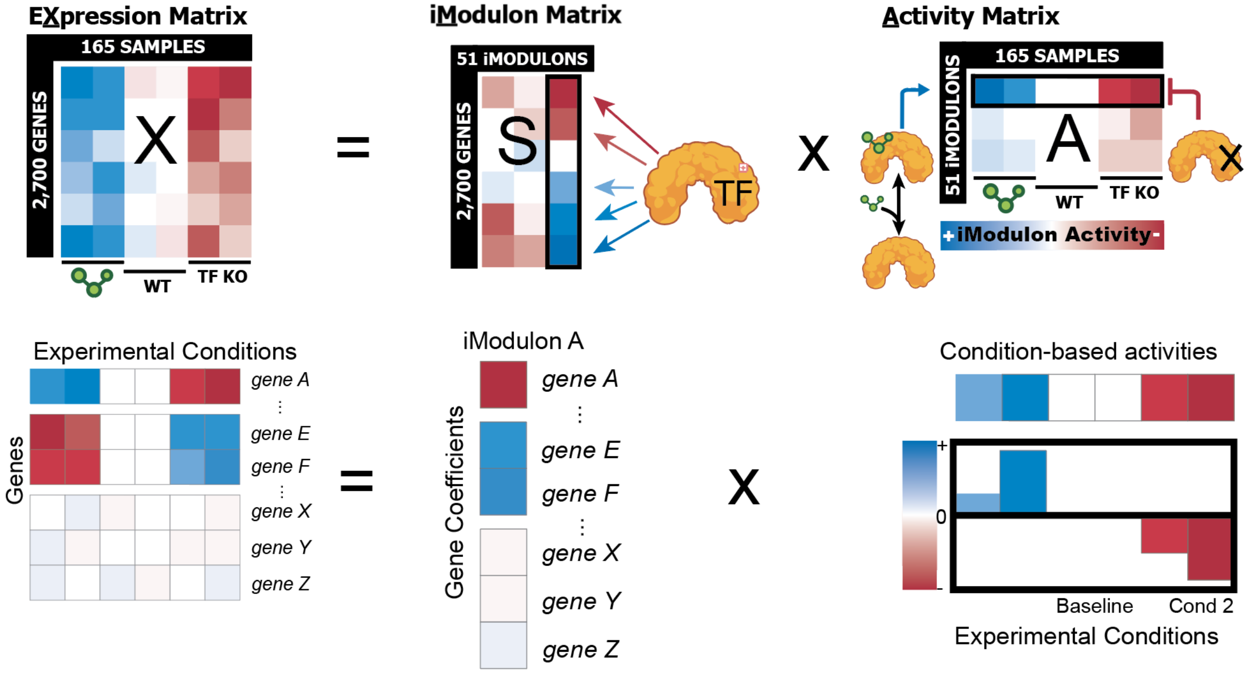


Figure 1.3: General structure of ICA decomposition. ICA decomposes the RNA-seq eXpression profile matrix (X) into two matrices: the iModulon (M) and Activity (A) matrices. The M matrix describes the effect of an underlying biological signal that induces transcriptional activation or repression. The A matrix describes the activity levels of the genes included in each iModulon. Each resulting independent component (column of M) contains a coefficient for each gene in the genome. These coefficients are then scaled by the condition-specific activities (row in A) to describe the contribution of each component to the expression or transcriptomic compendium. The summation of all iModulons and their activity will in return reconstruct the original X matrix. Both top and bottom panels convey a similar description of the main matrices involved in ICA.

1.5 Thesis Outline

In this thesis, we present the iModulon decomposition of *S. elongatus* in chapter 2 though (1) revealing iModulons with high regulon coverage to validate previous literature findings, (2) provide new functional insights to a few transcriptional regulators in *S. elongatus*, (3) propose putative transcription factors, and (4) provide activity-based analysis though clustering to study differentially expressed iModulons. Chapter 3 discusses the computational methodologies used to fulfill this project’s objectives.

Chapter 1 is currently being prepared for submission for publication of the material. **Tahani Al Bulushi**, Anand V Sastry, Kevin Rychel, Saugat Poudel, Reo Yoo, Siddharth Chauhan, Yuan Yuan, Cigdem Sancar, Richard Szubin, Bernhard Ø. Palsson, Susan Golden. (2021). “Machine learning reveals the transcriptional regulatory network and circadian dynamics of the cyanobacteria *Synechococcus elongatus* PCC 7942”. The thesis author is the primary author.

1.6 References

1. Abouzeed, Y. M., Baucheron, S., & Cloeckaert, A. (2008). ramR mutations involved in efflux-mediated multidrug resistance in Salmonella enterica serovar Typhimurium. *Antimicrobial Agents and Chemotherapy*, *52*(7), 2428–2434.

2. Alekshun, M. N., & Levy, S. B. (1997). Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. *Antimicrobial Agents and Chemotherapy*, *41*(10), 2067–2075.

3. Andersson, C. R., Tsinoremas, N. F., Shelton, J., Lebedeva, N. V., Yarrow, J., Min, H., & Golden, S. S. (2000). Application of bioluminescence to the study of circadian rhythms in cyanobacteria. *Methods in Enzymology*, *305*, 527–542.

4. Atsumi, S., Higashide, W., & Liao, J. C. (2009). Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nature Biotechnology*, *27*(12), 1177–1180.

5. Bailey, A. M., Ivens, A., Kingsley, R., Cottell, J. L., Wain, J., & Piddock, L. J. V. (2010). RamA, a member of the AraC/XylS family, influences both virulence and efflux in Salmonella enterica serovar Typhimurium. *Journal of Bacteriology*, *192*(6), 1607–1616.

6. 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. (2012). NCBI GEO: archive for functional genomics data sets—update. In *Nucleic Acids Research* (Vol. 41, Issue D1, pp. D991–D995). https://doi.org/10.1093/nar/gks1193

7. Brown, G. D., Yamada, S., & Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends in Neurosciences*, *24*(1), 54–63.

8. 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., … Sauer, U. (2012). Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. *Science*, *335*(6072), 1099–1103.

9. Choe, D., Lee, J. H., Yoo, M., Hwang, S., Sung, B. H., Cho, S., Palsson, B., Kim, S. C., & Cho, B.-K. (2019). Adaptive laboratory evolution of a genome-reduced Escherichia coli. In *Nature Communications* (Vol. 10, Issue 1). https://doi.org/10.1038/s41467-019-08888-6
10. Deng, M. D., & Coleman, J. R. (1999). Ethanol synthesis by genetic engineering in cyanobacteria. *Applied and Environmental Microbiology*, *65*(2), 523–528.

11. Diamond, S., Jun, D., Rubin, B. E., & Golden, S. S. (2015). The circadian oscillator in Synechococcus elongatus controls metabolite partitioning during diurnal growth. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(15), E1916–E1925.

12. Gottesman, S., & Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology*, *3*(12). https://doi.org/10.1101/cshperspect.a003798
13. Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent Component Analysis*. John Wiley

& Sons.

14. Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. *Annual Review of Microbiology*, *54*, 499–518.

15. Jones, S. A., Chowdhury, F. Z., Fabich, A. J., Anderson, A., Schreiner, D. M., House, A. L., Autieri, S. M., Leatham, M. P., Lins, J. J., Jorgensen, M., Cohen, P. S., & Conway, T. (2007). Respiration of Escherichia coli in the mouse intestine. *Infection and Immunity*, *75*(10), 4891–4899.

16. Kruse, O., & Hankamer, B. (2010). Microalgal hydrogen production. *Current Opinion in Biotechnology*, *21*(3), 238–243.

17. Marin, K., Suzuki, I., Yamaguchi, K., Ribbeck, K., Yamamoto, H., Kanesaki, Y., Hagemann, M., & Murata, N. (2003). Identification of histidine kinases that act as sensors in the perception of salt stress in Synechocystis sp. PCC 6803. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9061–9066.

18. McClure, R. S., Overall, C. C., McDermott, J. E., Hill, E. A., Markillie, L. M., McCue, L. A., Taylor, R. C., Ludwig, M., Bryant, D. A., & Beliaev, A. S. (2016). Network analysis of transcriptomics expands regulatory landscapes in Synechococcus sp. PCC 7002. *Nucleic Acids Research*, *44*(18), 8810–8825.

19. Minezaki, Y., Homma, K., & Nishikawa, K. (2005). Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *12*(5), 269–280.

20. Mohamed, E. T., Wang, S., Lennen, R. M., Herrgård, M. J., Simmons, B. A., Singer, S. W., & Feist, A. M. (2017). Generation of a platform strain for ionic liquid tolerance using adaptive laboratory evolution. *Microbial Cell Factories*, *16*(1), 204.
21. Nair, U., Thomas, C., & Golden, S. S. (2001). Functional Elements of the Strong psbAIPromoter of Synechococcus elongatus PCC 7942. *Journal of Bacteriology*, *183*(5), 1740–1747.

22. National Research Council, Division on Engineering and Physical Sciences, Board on Energy and Environmental Systems, Division on Earth and Life Studies, Board on Agriculture and Natural Resources, & Committee on the Sustainable Development of Algal Biofuels. (2013). *Sustainable Development of Algal Biofuels in the United States*. National Academies Press.
23. Nudler, E., & Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences*, *29*(1), 11–17.

24. Paithoonrangsarid, K., Shoumskaya, M. A., Kanesaki, Y., Satoh, S., Tabata, S., Los, D. A., Zinchenko, V. V., Hayashi, H., Tanticharoen, M., Suzuki, I., & Murata, N. (2004). Five histidine kinases perceive osmotic stress and regulate distinct sets of genes in Synechocystis. *The Journal of Biological Chemistry*, *279*(51), 53078–53086.

25. Poudel, S., Tsunemoto, H., Seif, Y., Sastry, A. V., Szubin, R., Xu, S., Machado, H., Olson, C. A., Anand, A., Pogliano, J., Nizet, V., & Palsson, B. O. (2020). Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. *Proceedings of the National Academy of Sciences of the United*

States of America, *117*(29), 17228–17239.

26. Raghavan, S., Manzanillo, P., Chan, K., Dovey, C., & Cox, J. S. (2008). Secreted transcription factor controls Mycobacterium tuberculosis virulence. *Nature*, *454*(7205), 717–721.

27. Rhee, H. S., & Pugh, B. F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 21*(1), Unit 21.24.

28. Ruffing, A. M., & Jones, H. D. T. (2012). Physiological effects of free fatty acid production in genetically engineered Synechococcus elongatus PCC 7942. *Biotechnology and Bioengineering*, *109*(9), 2190–2199.

29. Rychel, K., Sastry, A. V., & Palsson, B. O. (2020). Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. *Nature Communications*, *11*(1), 6338.

30. 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. (2018). A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. In *BMC Biology* (Vol. 16, Issue 1). https://doi.org/10.1186/s12915-018-0555-y

31. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A., & Palsson, B. O. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*, *10*(1), 5536.

32. Schirmer, A., Rude, M. A., Li, X., Popova, E., & del Cardayre, S. B. (2010). Microbial biosynthesis of alkanes. *Science*, *329*(5991), 559–562.

33. Shoumskaya, M. A., Paithoonrangsarid, K., Kanesaki, Y., Los, D. A., Zinchenko, V. V., Tanticharoen, M., Suzuki, I., & Murata, N. (2005). Identical Hik-Rre Systems Are Involved in Perception and Transduction of Salt Signals and Hyperosmotic Signals but Regulate the Expression of Individual Genes to Different Extents in Synechocystis. In *Journal of Biological Chemistry* (Vol. 280, Issue 22, pp. 21531–21538). https://doi.org/10.1074/jbc.m412174200

34. Singh, A. K., Elvitigala, T., Cameron, J. C., Ghosh, B. K., Bhattacharyya-Pakrasi, M., & Pakrasi, H. B. (2010). Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium. *BMC Systems Biology*, *4*(1), 105.

35. Suzuki, I., Kanesaki, Y., Mikami, K., Kanehisa, M., & Murata, N. (2001). Cold-regulated genes under control of the cold sensor Hik33 in Synechocystis. In *Molecular Microbiology* (Vol. 40, Issue 1, pp. 235–244). https://doi.org/10.1046/j.1365-2958.2001.02379.x
36. Tu, C.-J., Shrager, J., Burnap, R. L., Postier, B. L., & Grossman, A. R. (2004). Consequences of a deletion in dspA on transcript accumulation in Synechocystis sp. strain PCC6803. *Journal of Bacteriology*, *186*(12), 3889–3902.

37. Yang, Y., Feng, J., Li, T., Ge, F., & Zhao, J. (2015). CyanOmics: an integrated database of omics for the model cyanobacterium Synechococcus sp. PCC 7002. *Database: The Journal*

of Biological Databases and Curation, *2015*. https://doi.org/10.1093/database/bau127

38. Yanofsky, C. (1988). Transcription attenuation. *The Journal of Biological Chemistry*, *263*(2), 609–612.

39. Zalucki, Y. M., Dhulipala, V., & Shafer, W. M. (2012). Dueling regulatory properties of a transcriptional activator (MtrA) and repressor (MtrR) that control efflux pump gene expression in Neisseria gonorrhoeae. *mBio*, *3*(6), e00446–12.

Chapter 2

Results

2.1 The *S. elongatus* transcriptome consists of regulatory and functional iModulons

To discover the transcriptional regulatory landscape of *S. elongatus*, we extracted 323 publicly available RNA-seq expression profiles from 12 individual studies, see **Table 1**, (Fleming & O'Shea, 2018; Markson et al., 2013; Piechura et al., 2017; Puszynska & O'Shea, 2017a, 2017b; Ruffing, 2013) as found in NCBI's Sequence Read Archive (SRA). Each dataset was processed using standardized RNA-seq and quality control protocols to produce a final compendium consisting of 262 expression profiles for *S. elongatus* (see chapter 3 for Methods). ICA of this compendium resulted in 51 iModulons, each of which describes a gene cluster that is expressed independently from other clusters across all conditions in the compendium (Sastry et al., 2019). The discovered 51 iModulons capture 41% of the genes in the genome and explain 73% of the variance in the transcriptome. 1108 unique genes were enriched in the iModulons, and 305 appeared in more than a single iModulon, indicating the presence of multiple controlling regulators.

While a few iModulons recapitulate predefined regulon structures, other iModulons might capture a portion of a given regulon or show no regulon enrichment. This implies that the regulon-iModulon relationship is complementary and can be described using two measures (Rychel et al., 2020): iModulon recall (MR) and regulon recall (RR), referring to the fraction of shared genes given iModulon genes and regulon genes, respectively (**Figure 2.1a**). iModulons are not only interpreted through regulon enrichment, but can also be characterized through statistical enrichments between their constituent genes and other knowledge tools such as KEGG pathway, GO annotations, or other knowledge types found in the literature (Kanehisa et al., 2021; The Gene Ontology Consortium & The Gene Ontology Consortium, 2019). In total, we categorized the 51 iModulons into five main categories: Regulatory, Functional, Genomic, Uncharacterized, and Single Gene (**Figure 2.1b**). ‘Regulatory’ iModulons are enriched in a transcriptional regulator, whereas ‘Functional’ iModulons are related to a particular biological function. ‘Genomic’ iModulons represent the effect of genomic changes, such as gene knock-outs. ‘Single Gene’ iModulons track the expression of a single gene and are likely decomposition artifacts from ICA (McConn et al.). Finally, ‘Uncharacterized’ iModulons are those that do not fall into any of the previous categories, likely due to the presence of many uncharacterized genes. We also further characterized these categories to obtain a deeper understanding of their systems-level metabolic roles (**Figure 2.1c**). iModulons can also be ranked based on their explained variance in the data, which helps to construct a hierarchical understanding of their regulatory contributions to the adaptive response machinery in *S. elongatus* (**Figure 2.2**).

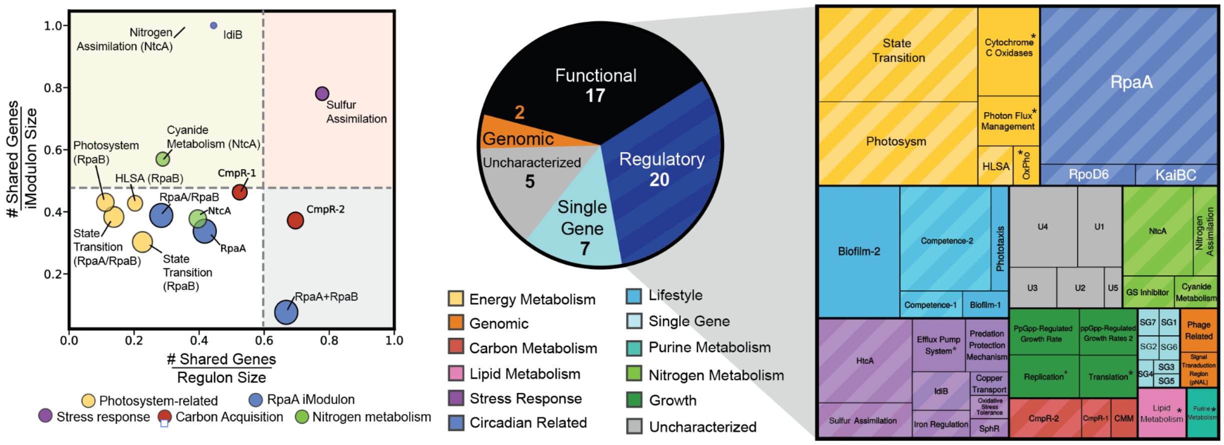


Figure 2.1: iModulon and regulon relationship and iModulon categories. (a) Regulon enriched iModulons are those that recapitulate or show enrichment to well-defined regulons. iModulon recall (MR) and regulation recall (RR) are two measures used to describe enrichment overlap. The size of the scatter dots is indicative of iModulon size (i.e. number of genes contained in the iModulon) and the color is indicative of its corresponding function. The top-right quadrant (highlighted in light orange) means well-matched enrichments with high MR and RR. **(b)** Pie chart dividing iModulons into five major categories deepening on their function and/or properties: Functional, Regulatory, Genomic, Uncharacterized, and Single Gene. **(c)** A treemap illustrating subcategories of the main five categories for each iModulons. Functional iModulons are marked with asterisks (“*”) after iModulon name (ex. Cytochrome C Oxidases), and Regulatory iModulons are distinguished with a diagonal pattern (ex. RpaA). Legends for the treemap are included under the pie chart.

Apart from the iModulon matrix (**M**), the other output of ICA is the activity matrix (**A**) which enables condition-dependent comparison studies across multiple iModulons (**Figure 1.3**), demonstrating the tractability of iModulons as compared to traditional differential expression analysis (DEGs) since the number of iModulons is significantly fewer than the number of genes in the organism's genome. This feature facilitated quantifiable comparisons of iModulon activities between contrasting data points, such as light versus dark exposures, wildtype versus genetic perturbation, and adverse environmental stressors. iModulon activities can also be clustered to identify similar iModulons that compose core biological phenomenons. To mitigate batch effects between the expression profiles gathered from the six independent datasets, activity levels were centered to a reference condition within each dataset. **Table 1** lists all samples extracted from NCBI's SRA with a description of each study and other metadata.

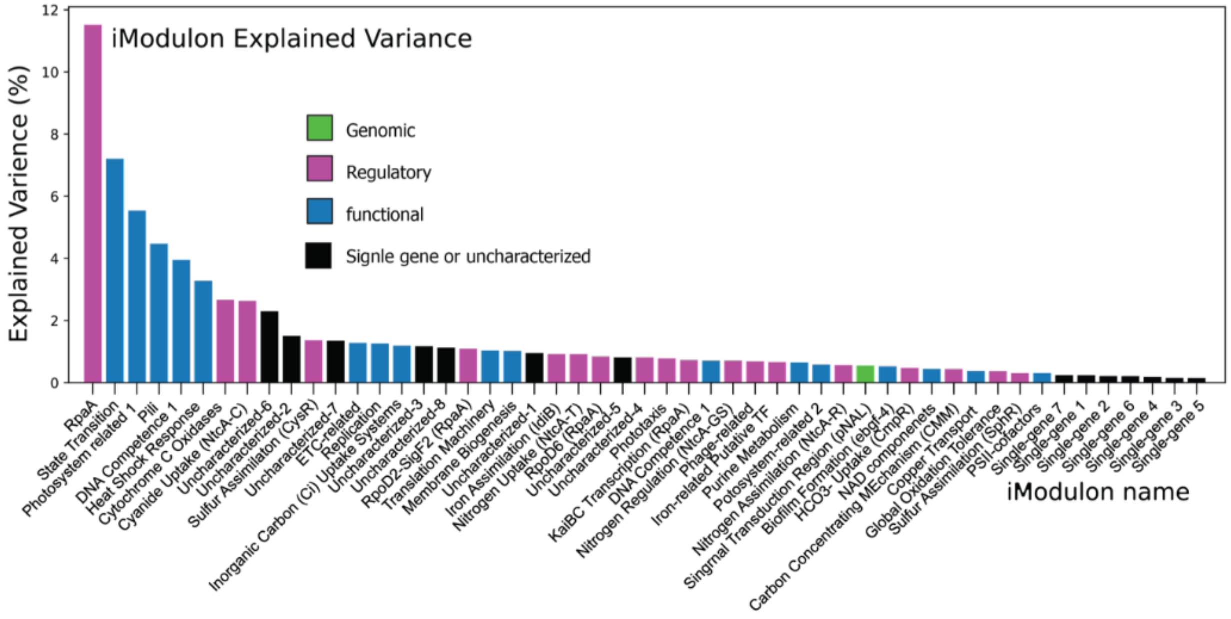


Figure 2.2: iModulon explained variance distribution. iModulons that explain a large portion of the data are those that are regulated by the endogenous clock or are responsible for major metabolic shifts and expression changes. Together, all iModulons explain 73% of the variance in the transcriptomic data. The legend shows iModulon categories (Regulatory, Functional, Genomic, and Uncharacterized).

The ICA framework also enables simultaneous analysis of TRNs at the gene and genomic scales by quantitatively capturing complex regulatory behaviors describing coregulation of a single gene by multiple regulators, regulation of multiple genes by the same regulator, and coordinated expression levels of multiple iModulons under various conditions. We will demonstrate these complex regulatory cases when discussing individual iModulon in this section. Another attractive quality of ICA is its ability to compute condition-dependent activity levels, indicative of the underlying transcription factory activity, for each iModulon across every sample in the compendium. This enables efficient condition-dependent comparison studies across multiple iModulons or samples of interest, giving it an advantage over differentially expressed genes studies since the number of genes in a given organism is larger than the number of iModulons. The reported activity levels are centered on a reference condition for each collected RNA-seq bioproject. A detailed section is dedicated to elucidating the different activity-based analyses enabled through the use of ICA.

Table 1. Original studies of all RNA-seq data extracted from NCBI's SRA repository

BioProject	Project name	#n	GEO Accessions	Study Description	Reference Condition
PRJNA140271	RNA	3	GSE29264	Global transcriptome architecture of *S. elongatus* PCC 7942 (Vijayan et al., 2011)	wt
PRJNA354335	RpaA	24	GSE89999	Gene expression analysis in clock rescue and RpaA mutant clock rescue strains during dark and light (Puszynska & O'Shea, 2017a).	wt_dusk_000
PRJNA196229†	FFA	17	GSE45762	RNA-seq analysis of targeted mutagenesis to improve the secretion of free fatty acids in engineered S. elongatus strains (Ruffing, 2013)	wt_100h
PRJNA372989	S72	19	N/A	Analysis of changes in salinity, temperature and pH levels on *S. elongatus* to address different stress acclimation (Billis et al., 2014)	ref
PRJNA401742† PRJNA404081† PRJNA403840† PRJNA415380†	ppGpp	94	GSE45762 GSE103463 GSE103704 GSE103644 GSE105774	Analysis of gene expression between wild-type and rel- strains under constant light to study the role of ppGpp in *S. elongatus* (Puszynska & O'Shea, 2017b).	0wt_dawn_000
PRJNA412032†	Light	60	GSE104203	Measurement of genome-wide gene expression grown under simulated natural light conditions (Piechura et al., 2017).	CD_0.5h
PRJNA221220†	Clock	18	GSE51112	Understanding the extent of RpaA in controlling circadian gene expression (Markson et al., 2013)	WT_24h
PRJNA472248†	Sigma Factors	72	GSE114693	A deep analysis of RpaA-dependent sigma factor cascade in S. elongatus (Fleming & O'Shea, 2018)	del_rpoD2_0h
PRJNA506580	H_2O_2	4	GSE122841	Compare hydrogen peroxide stress tolerance through gene expression of wild type and OsTPX-expressing *S. elongatus* under the normal and stress conditions (Kim et al., 2018)	wt_normal
PRJNA588336	DHAR	4	GSE140121	Expression of heterologous OsSHAR gene improved glutathione-dependent antioxidant system and redox balance in *S. elongatus*	wt_normal
PRJNA642094	RNA-2	1	N/A	RNAseq Co-culture comparison of two species	N/A
N/A	RNA-3	1	N/A	No trace of the original study	N/A
Total		**317**			

† Bioprojects that passed the quality control (QC) pipeline and are included in the ICA analysis.

All results generated from the analysis of this study are deposited in the form of interactive dashboards and are included in iModulonDB for *S. elongatus* (see chapter 3 for Methods). The presented TRN covers all available public RNA-seq datasets gathered as of August 20, 2020. We encourage researchers to use the website to refer to the discovered iModulons, including their regulators and gene sets, in this study.

In this section, we discuss three regulon-enriched iModulons to confirm the validity of our TRN composition, the discovery of hypothetically new regulons and their associated TFs, the functional insights that might expand current knowledge concerning different metabolic and ecological processes in *S. elongatus*, and the annotation of functionally-obscure iModulons through differential activity analysis.

2.2 iModulons capture predefined transcriptional regulators

We first describe two iModulons whose structures significantly overlap with well-characterized regulons from the literature: RpaA and CmpR. We evaluate the activity levels of the RpaA iModulon to show that its regulator activity is in agreement with existing knowledge and provide additional insight into the CmpR regulon.

RpaA, the regulator of phycobilisome association A, is a master transcription factor that binds to 170 downstream gene targets (Markson et al., 2013). 68 of these gene targets were captured in an iModulon, which led us to name it the RpaA iModulon (**Figure 2.3a**). The incompleteness of the RpaA iModulon is likely a result of other confounding regulators affecting the remaining genes in the RpaA regulon, or maybe because iModulons sometimes identify multiple dynamic subsets of a single regulon (Lamoureux et al., 2021; Sastry et al., 2020). The RpaA iModulon was most active in the moments leading to and during dusk (PRJNA:412032), which is consistent with previous findings for its constituent genes (dusk peaking class I genes) (Diamond et al., 2015; Markson et al., 2013) (**Figure 2.3b**). In addition, *rpaA* null-mutants exhibited negative iModulon activity, indicating under-expression of genes in the iModulon as compared to wildtype samples during dark conditions (PRJNA:412032). This clearly confirms that the RpaA iModulon behaves similarly to the expected RpaA TF.

The TF CmpR regulates the acquisition of inorganic carbon in response to cellular CO_2 levels. The regulon contains the transcripts of the bicarbonate transport system (*cmpABCD*) and its own gene (Omata et al., 1999, 2001). Speculations regarding a repressive role for CmpR have been made to influence the expression of BicA (some *mnh* genes), SbtA, NDH-I_3 (*ndhF3-ndhD3-cupA-cupS*), and NDH-I_4 (*0307-ndhF4-cupB and ndhD4)* under high CO_2 conditions(Pan et al., 2016) (**Figure 2.3c**). We identified two iModulons with significant overlap with each other and the CmpR regulon and (**Figure 2.3d a-f**). The CmpR-1 iModulon contained 15 genes, including both *cmpABCD and ndhF3-ndhD3-cupA-cupS* operons. The expression of the latter operon encodes the high-affinity CO_2 uptake system NDH-I_3 and is induced in low-CO_2 conditions (Maeda et al., 2002; Ohkawa, 1998). The CmpR-2 iModulon captured 14 genes regulated by CmpR, such as *sbtA* encoding a sodium-dependent bicarbonate uptake system (Shibata et al., 2002), *mnhCD* encoding the NDH-1 complex-associated Na+/H+ antiporter (Price, 2011), and the *cmpR* gene. Based on the regulon coverage for the CmpR-2 iMdoulon, we believe an additional TF jointly regulates the CmpR-2 iModulon. This dual-regulation might explain the presence of *ndh* genes, encoding NDH-1 complex, and confirms the observed biological connection between respiration to carbon acquisition is *S. elongatus* (Battchikova et al., 2011). We examined both iModulon activities under high light pulse and shade pulse conditions, considering the direct implications of light intensities and CO_2 levels on the activity of CmpR (Pan et al., 2016; Woodger et al., 2003). CmpR-1 was consistently down-regulated because of adequate supplementation of carbon sources in the original study. CmpR-2 showed expected minimal expression under sufficient CO_2 and high light pulse conditions (Pan et al., 2016). During the shade pulse, however, a reduction was observed. The decrease in light intensity in the shade pulse condition might reduce the cellular demand for CO2, since a reduction in photonic energy, causing a reduction in the amount of electrons, ATP, and eventually the need to import CO_2. Together, both CmpR-1 and CmpR-2 iModulons demonstrate the characteristics of the CmpR TF during CO_2 sufficient conditions and suggest an additional regulation in the CmpR-2 iModulon. The complete activity for both iModulon across all samples in the transcriptomic compendium is provided in the supplemental **Figure A.2**.

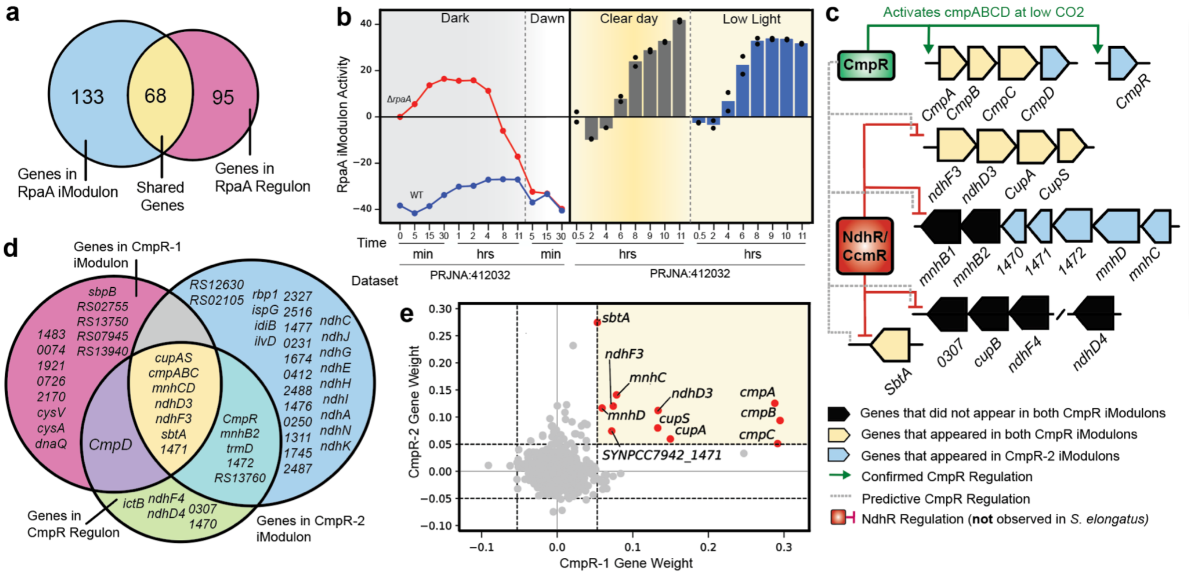


Figure 2.3: iModulons validate retrospective literature observations. (a) Venn diagram between the RpaA regulon and RpaA iModulon. **(b)** Activity of the RpaA iModulon genes between wildtype and rpaA deletion mutants (left), and wildtype samples during clearday condition and constant Low Light conditions (right). **(c)** General understanding of CmpR and its homologs (NdhR and CcmR) binding sites and regulations. Note that the red solid arrows corresponding to NdhR/CcmpR are not observed in *S. elongatus.* However, based on recent suggestions (Pan et al., 2016), CmpR behaves as a repressor similar to the repressive role of NdhR/CcmpR in *S. elongatus* (gray dashed lines). **(d)** three-way Venn diagram showing the gene distribution between CmpR-1 and CmpR-2 iModulons and CmpR regulon. **(e)** Scatter plot showing the shared genes (red) between both CmpR-1 and CmpR-2 iModulons.

2.3 iModulons generate hypotheses by elucidating new transcription factors and their function

After confirming that iModulons are representative of the *S. elongatus* TRN, we analyzed the remaining iModulons to either discover new regulatory patterns or expand the characterization of current regulons. We illustrate this by elucidating the configuration and activity levels of iModulons related to various biological and metabolic categories. We, first, present putative transcription factors and their genetic targets followed by an extensive discussion of functional iModulons and, if applicable, their association to circadian clock dynamics.

2.3.1 Putative TF may coordinate Fur-IdiB crosstalk in iron-deprived conditions

Iron homeostasis has been shown to prevent oxidative stresses and maintain photosynthesis in cyanobacteria (Kranzler et al., 2013). This is predominantly because iron serves as a cofactor for membrane-bound protein complexes and other mobile electron carriers within the photosynthetic apparatus (Cheng & He, 2020). Similar to other photosynthetic organisms, *S. elongatus* has evolved regulatory molecular switches that stringently regulate iron acquisition and metabolism, of which Fur and IdiB are the most understood to date. Fur is a ferric uptake regulator and IdiB is an iron deficiency-induced protein B (Ghassemian & Straus, 1996). We identified the idiB iModulon that showed a significant overlap to the IdiB regulon (**Figure 2.4a**). Past studies have reported that Fur and IdiB are connected in their activation under the influence of reactive oxidative species (ROS), suggesting an existing cross-talk between both TF (Nodop et al., 2008; Yousef et al., 2003). However, it also has been demonstrated that *idiB* does not contain a Fur-box consensus sequence upstream of its coding region, an indication that Fur does not regulate the expression of *idiB* (Nodop et al., 2008; Yousef et al., 2003). Despite this, Yousef et. al. proposed that an existing unknown global repressor that is sensitive to the presence of ROS binds upstream of *idiB and idiC*, and could elucidate the observed crosstalk between IdiB and the Fur regulon. Although we have not identified a Fur iModulon in this study, we found an iron-related iModulon that regulated *idiB* gene, and subsequently, the IdiB iModulon. This is an example of a nested iModulon structure where one iModulon contains a gene that encodes a TF of another iModulon. In addition to containing *idiB*, the iron-related iModulon also captures *idiC,* iron transporter genes *futA2BC* whose regulator has not been identified prior to this study (Nodop et al., 2008)*,* and additional hypothetical genes, one of which has a putative iron uptake function (*synpcc7942_2169*). We hypothesize that this iron-related iModulon might be regulated by the product of *synpcc7942_2170* because it contains a helix-turn-helix DNA binding domain. To support our hypothesis further, this iModulon also regulates *isiA* and *isiB* (genes belonging to the Fur regulon) that are normally expressed under iron-deficient conditions (**Figure 2.4b**). This observation strongly suggests that *synpcc7942_*2170 might be the global regulator that connects the activity of Fur and IdiB (**Figure 2.4c**). Not much is known about *synpcc7942_2170* and the protein it encodes, which warrants future studies to investigate its involvement in the structural interplay of iron regulation in *S. elongatus.*

To understand the regulatory effect the iron-related iModulon exerts on the IdiB iModulon, we compared their activities and observed that their expression activity varied the most in PRJNA:412032 (**Figure A.2**) (Piechura et al., 2017). During natural light conditions, the iron-related iModulon appeared to be leveled off followed by a sharp increase before dusk, while the IdiB iModulon remained downregulated (**Figure 2.4d**). From this observation, we developed two possible interpretations: the iron-related iModulon could result in the downregulation of IdiB and its genetic targets during daytime. Secondly, if *synpcc7942_2170* is responsive to accumulating ROS*,* the transient increase in activity of the iron-related iModulon during natural light conditions can be viewed as part of the preparation process for the nighttime redox restoration (Welkie et al., 2019).

Similar to the iron-related iModulon, we identified three autoregulated iModulons with genes containing the helix-turn-helix DNA binding domain, evidencing that their protein products could regulate the remaining genes within the iModulon structure. To support our hypotheses regarding these iModulons, we performed a comparative analysis between *S. elongatus* iModulons and iModulons from previously published microbes (Poudel et al., 2020; Rychel et al., 2020; Sastry et al., 2019). We used reciprocal BLAST hits to generate one-to-one orthology between *S. elongatus* and these organisms to obtain orthologous gene pairs. Once determined, we applied a distance metric to identify homology between the three *S. elongatus* iModulons to iModulons from other organisms. From this analysis, we discovered that the putative TF for each of the *S. elongatus* autoregulated iModulons have conserved genetic sequences to TFs regulating iModulons from other organisms.

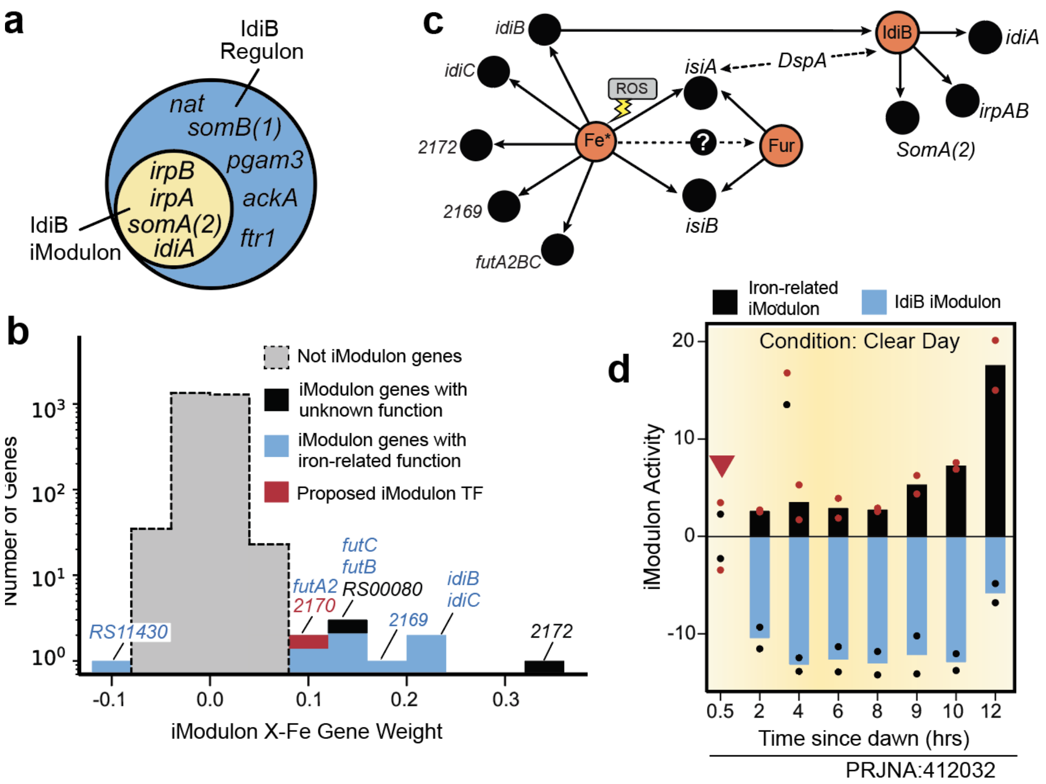


Figure 2.4: iModulons Discover New Potential Transcription Factors (a) Venn diagram showing the shared genes between the IdiB regulon and the ICA-computed idiB iModulon **(b)** Gene weight plot for the putative iron-related iModulon regulated by the HTH DNA-binding protein *synpcc7942_2170* (2170 for short). The legend shows the category of genes in this iModulon **(c)** Redefined regulatory cross-talk between all iron inducing transcription factors (Fur, IdiB and the proposed 2170), a continuation of the communicated regulatory network in Yousef et. al. (Yousef et al. 2003) **(d)** iModulon activity comparison between the IdiB iModulon and the putative 2170 iModulon. Reference condition is shown with a triangle and dots represent biological replicates for each condition in the PRJNA:412032 dataset.

2.3.2 CysR and sulfur assimilation

CysR is a transcriptional regulator belonging to the Crp-family of prokaryotic regulator proteins that facilitates the acclimation of *S. elongatus* to conditions of low sulfur. Inactivation of the *cysR* gene prevents the increase in activity of sulfate premeases and sulfur assimilation into the cytosol under sulfur limiting conditions (Nicholson et al., 1995). However, CysR is not essential for growth when abundant levels of sulfate or thiosulfate is provided in the culture medium (Laudenbach & Grossman, 1991). Instead, CysR is involved in regulating the growth of sulfur-containing compounds since *cysR* mutants fail to utilize sulfur-resources in the media. We identified CysR in the Sulfur Assimilation iModulon, which also includes ABC transporter preseamses (*cysT, cysW, CysU,* and *CysV*) and sulfate-binding genes (*cysA, cysP, sbpA, sbpB)*. Other captured genes include *rdhA* (rhodanese-like protein), *sir* (Sulfite reductase, ferredoxin dependent) and *per1* (1-Cys peroxiredoxin). Additionally, a plasmid gene of hypothetically conserved function (anL43) was also captured which suggests that it might be related to sulfur metabolism in *S. elongatus*. From a cross iModulon correlation test, the CysB iModulon in *E. coli,* in which CysB regulates sulfur uptake during sulfur deficiency, was correlated to this iModulon with Pearson R correlation of 0.6 (**Figure 2.5a**). However, no sequence homology was detected between the genes *cysB* and *cysR*. Further investigation of this iModulon is encouraged.

2.3.3 HrcA might regulate heat shock response in S. elongatus

A set of chaperon and heat shock proteins comprise the heat shock resistance (HSR) iModulon is *S. elongatus* and showed a significant correlation and nearly identical orthologs in the *Mycoplasma pneumoniae* HrcA iModulon (Pearson R 0.87), indicating that these genes are modulated in similar ratios across the two organisms (**Figure 2.5b**). *hrcA* (*synpcc7942_RS03160*) in *S. elongatus* encodes a CIRCE-specific transcriptional repressor of the heat shock genes and is highly conserved in genomes of cyanobacteria (Nakamoto & Kojima, 2017). A study determined that the mRNA level of *groESL1* increased in the hrcA mutant of *S. elongatus* during heat-treatment at 30 ^{0}C followed by a reduction in level of mRNA accumulation at 60-mins post heat-treatment (Saito et al., 2020).

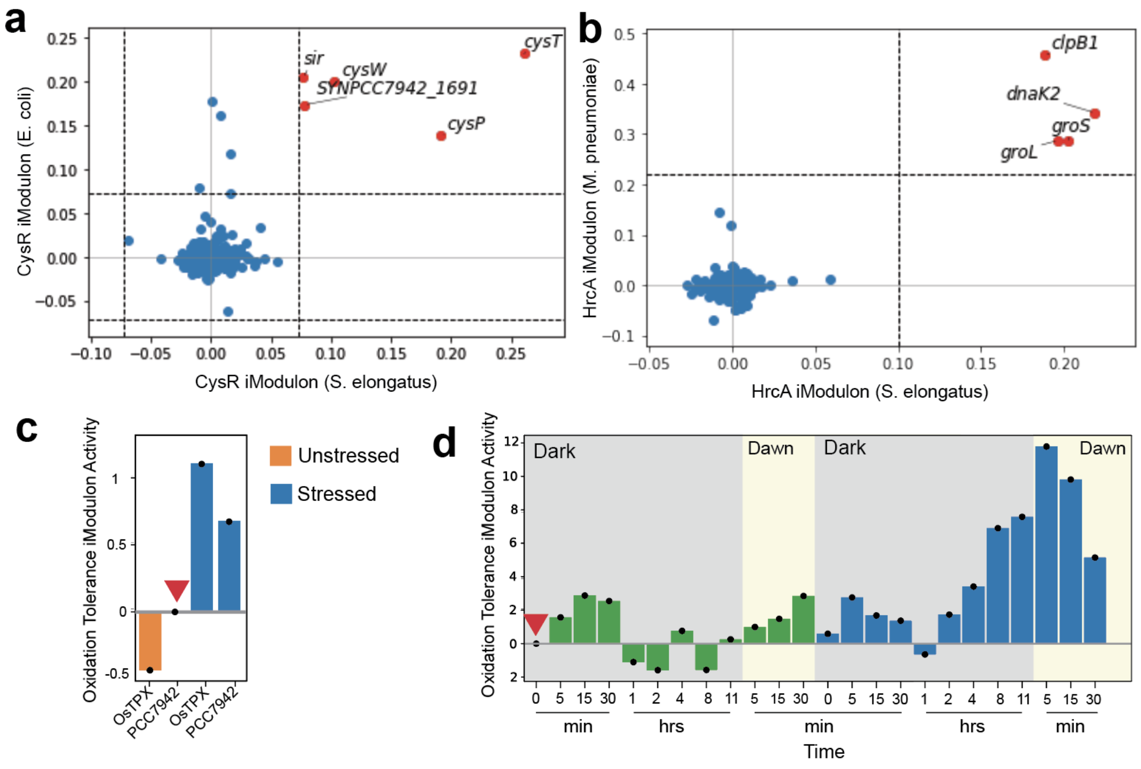


Figure 2.5: Putative regulation patterns and new transcription factors (a-b) Scatter plots comparing gene weights of iModulons found in different datasets.Genes shown in red are members in both iModulons and horizontal and vertical dashed lines indicate iModulon thresholds. **(a)** Comparison of the *S. elongatus* CysR iModulon to the *E. coli* CysB iModulon. **(a)** Comparison of the *S. elongatus* HrcA iModulon to the*M. pneumoniae* HrcA iModulon. **(c)** stressed vs unstressed activity of the oxidative stress tolerance iModulon from PRJNA:506580. Labels to the right indicate the applied condition (i.e. stressed with the addition of H_2O_2 and unstressed without the addition of H_2O_2). **(d)** oxidative stress tolerance iModulon activity comparison between WT and rpaA mutant strains. Reference condition is marked with a red triangle in panel c and d.

2.3.4 Activation of antioxidants by synpcc7942_0110

The oxidative tolerance iModulon is comprised of antioxidants and reductases that are directly involved in mitigating superoxides and also act as electron carriers in many biochemical processes. These genes include *trxB* (thioredoxin reductase), *rbr* (rubrerythrin), and *tpxA* (2-Cys peroxiredoxin), *parB* (cellular oxidant detoxification nuclease), *synpcc7942_0109* (Ferritin-like protein), *synpcc7942_1648 (putative ferric uptake regulator)*, *synpcc7942_RS03965* (hypothetical protein), *synpcc7942_0110 (HTH-DNA binding protein)*, and *sigF2*. Thioredoxins like TrxB have been shown to participate in oxidative response and particularly respond to high light stress and chlorophyll formation (Pérez-Pérez et al., 2009). Rubrerythrin proteins function as peroxidases and electron carriers to provide oxidative stress protection from metabolically generated hydrogen peroxides (Sztukowska et al., 2002). TpxA has signaling importance to maintain hydrogen peroxide scavenging (Toledano & Huang, 2016). Since oxidative damage can damage genetic material, ParB encodes a nuclease that is involved in DNA repair. Most importantly, we were interested in *synpcc7942_0110* since it could be the protein regulating the expression of this iModulon. The upstream gene of *synpcc7942_0110* (*synpcc7942_0109*) encodes an oxidative damage protein. Moreover, the activity of this iModulon was greater in stressed conditions compared to non-stressed conditions (Kim et al., 2018) (PRJNA:506580, **Figure 2.5c**) and showed an upregulated in the rpaA mutant samples, suggesting that RpaA exerts a regulatory influence on synpcc7942_0110 protein (**Figure 2.5d**). Further experimental validation is recommended to further test the predictions discovered by the ICA framework through gene knockout experiments of the aforementioned putative transcription factors.

2.4 Functional iModulons provide additional knowledge regarding their overarching regulators

iModulons result from a top-down analysis of large transcriptomic data that efficiently captures the TRN configuration. Since *S. elongatus* lacks a completely defined regulon structure, many of the resulting iModulons were strongly associated with biological processes that are likely to be controlled by undiscovered regulators. In this section we target three biological and metabolic systems, whose relationship with regulons are yet to be defined. We first elucidate the regulatory patterns of RpaB under dynamic light conditions, discuss the modes of regulation of NtcA, and expand on the genetic targets that are associated with pili-related mechanisms.

2.4.1 The role of RpaB in regulating light-dependent genes during dynamic light changes

Given the important role of photosynthesis in powering metabolic pathways, we were interested in studying the implications of light fluctuations on the transcription of the photosynthetic apparatus and its intermediates. In order to study such effects, we utilized data obtained from a study where four dynamic light conditions were applied (PRJNA:412032) (Piechura et al., 2017). The two base conditions were 'Clear Day', resembling natural daylight in a parabolic manner, and continuous 'Low Light' leveled at low photon rates (**Figure 2.6**). Cells were also exposed to abrupt increases ('High Light pulse') and decrease ('Shade pulse') in light intensities for 1 hour between hours 8-9. Note that the Low Light condition is not to be confused with the Constant Light condition (often denoted as LL in the literature). We will be referring to these conditions when discussing three light-dependent iModulons to study the regulatory role of RpaB: Photosystem, High Light Stress Acclimation, and State Transition.

A major regulator that controls light-dependent genes in *S. elongatus* is RpaB, a paralog of RpaA that binds to some circadian gene promoters (Hanaoka et al., 2012; López-Redondo et al., 2010). RpaB is a repressor of the high light-inducible (*hliA* and *rpoD3*) genes during standard (non-stressed or LL) conditions by binding to the high light regulatory 1 (HLR1) sequence (Seki et al., 2007). However, during high light stress, the phosphorylation state RpaB is altered to de-represses the high light-inducible genes (López-Redondo et al., 2010; Moronta-Barrios et al., 2012). Furthermore, RpaB is also an activator for the Photosystem (PS) I genes, containing HLR1, during standard conditions to maintain viability (Moronta-Barrios et al., 2012; Seino et al., 2009). Thus, RpaB works for positive regulation of PSI genes and negative regulation for the high light-inducible genes during standard and high light stress conditions, respectively.

In this context, the Photosystem iModulon contained two anti-correlated gene clusters. The positively weighted gene cluster contained phycobilisomes and reaction centers for both PSI and PSII. The negative gene cluster encompassed high light-inducible genes and *nblA*, a major regulatory factor in the bleaching process and the degradation of phycobilisomes. Of these systems, only PSII components are not reported to be regulated by RpaB (Kato et al., 2011; Seino et al., 2009). The negative gene cluster is consistent with their function, considering their activation only during high light stress conditions.

To understand the regulatory behavior of RpaB on the Photosystem iModulon, we examined the activity levels of this iModulon across the aforementioned light conditions (Clear Day, Low Light, High Light Pulse, and Shade Pulse). For all conditions, expression levels were in agreement with the reported results of the original study, implying that the transcription of PSI and PSII follow a general trend of class I genes (**Figure 2.6-c**). The upregulation in Shade Pulse is ascribed to RpaB repression of PSI genes during photoreduction while the downregulation in High Light Pulse is due to the activation of the high light-inducible genes. The latter is important to induce the expression of high light-inducible genes to mitigate the production of reactive oxidative species as a direct result of High Light Pulse. Therefore, the Photosystem iModulon demonstrates two anti-correlated systems that are tightly regulated by RpaB. We will refer to the Photosystem iModulon as a baseline reference when discussing the following two RpaB-regulated iModulons.

The High Light Stress Acclimation (or HLSA) iModulon represents the photosystem protection mechanism against high light stress. Expression levels opposite to the direction of pulses were observed (**Figure 2.6-d**). The HLSA iModulon captured *hliA* and *rpoD3*, with the highest gene weights, which were negatively weighted in the baseline Photosystem iModulon. Unlike PSI subunits and protochlorophyllide reductases, the genetic components of ferredoxins, photolyases, quinone binding, and PSII reaction centers remained highly weighted in the HLSA iModulon. We theorize that these genes play vital roles in acclimating to high light intensities (Seino et al., 2009). Other genes were also present in this iModulon (Supplementary Figure X). We also note that both RpaB and SigF2 transcripts are negatively weighted during high light acclimation, consistent with previous findings (Moronta-Barrios et al., 2012).

The last iModulon describes state transition, a rapid physiological adaptation mechanism that involves the redistribution of absorbed excitation energy between PSI and PSII (Mullineaux & Emlyn-Jones, 2005). A conserved cyanobacteria gene, *rpaC* for Regulator of Phycobilisome Association C, in the State Transition iModulon was previously shown to optimize the utilization of absorbed photons under limiting light conditions for growth (Emlyn-Jones et al., 1999). Further suggestions regard a second role for RpaC as a protector factor against photoinhibition, by which it curtails the rate of PSII damage through antenna size reduction (G. Finazzi et al., 2001; Giovanni Finazzi & Forti, 2004).

In light of this, we examined the activity of the State Transition iModulon and compared it to the activity of the baseline Photosystem iModulon (**Figure 2.6-c**). Clear Day samples showed a higher iModulon activity just before dusk, where light intensity diminishes, with the highest peak at CT =12 hr. No significant changes were observed in Low Light. This is because the difference in light intensities when shifting from Low Light to dusk is not drastic enough to induce high *rpaC* expression. Expectedly, a dramatic upregulation in Shade Pulse was observed due to the significant difference between Clear Day and Shade Pulse. Furthermore, rapid downregulation in High Light Pulse was observed. We interpret this result to the activation of state 1 transition, in which the absorbed excitation energy is diverted from PSI to PSII, since PSI predominantly absorbs low photon intensity (Mullineaux & Emlyn-Jones, 2005) Similarly, in the case of the high light pulse, PSII highly receives most excitation energy which then induces state 2 traditions to spill some of the excited energy to PSI causing a reduction in *rpaC* expression. Other genes within this iModulon cluster, such as phosphoenolpyruvate synthase (*ppsA*), indicate an upregulation of gluconeogenesis in preparation of metabolic partitioning that usually occurs at subjective dusk (or in this case during Shade Pulse since it mimics dusk condition) in *S. elongatus*. We also note that the State Transition iModulon generated positive activity levels during dark time, considering a few dark-inducing genes that are also included in its configuration (**Figure A.3b**).

Based on the presented results, we theorize that the State Transition iModulon could be another candidate target for RpaB. Researchers are encouraged to use the gene cluster in this iModulon to better understand the global transcriptional regulation of RpaB.

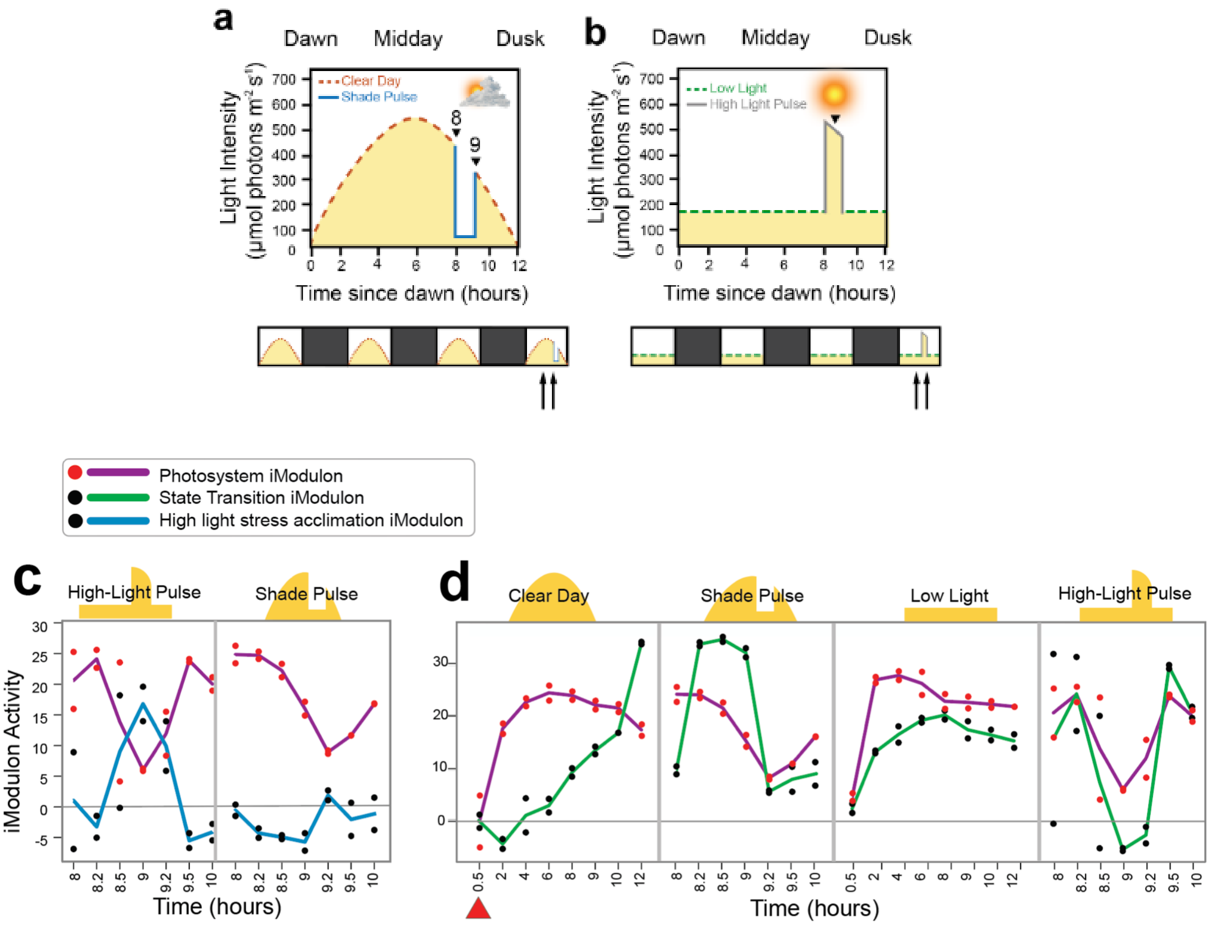


Figure 2.6: RpaB-related iModulon activities and light intensity profiles. (a) Experimental setup of Clear Day conditions with maximum photon intensity of 600 μmol photons m^{-2} s^{-1}. Shade Pulse exposure applied at 8 hour for a duration of one hour after dawn during the fourth light period for one hour before being returned to Clear Day conditions. **(b)** Experimental setup of Low Light conditions constant at 50 μmol photons m^{-2} s^{-1} throughout the day. High Light Pulse exposure applied at 8 hour for a duration of one hour after dawn during the fourth light period for one hour before being returned to Low Light conditions. **(c)** Activity comparison between Photosystem (PS) iModulon and high-light acclimation iModulon, both of which are regulated by RpaB. The plots show their activities under High-Light and Shade Pulse conditions. **(d)** Activity comparison between Photosystem (PS) iModulon and Low-light acclimation iModulon across all four light conditions. Legends for panels c and d are shown above panel c. Reference condition is shown with a red triangle, dots represent biological replicates and lines represent average expression between biological replicates.

2.4.2 Different Nitrogen Sources Lead to Distinct Transcriptional Modes of NtcA

The regulation of nitrogen metabolism in *S. elongatus* is mainly operated by the TF NtcA (Forchhammer & Selim, 2020), which induces nitrogen assimilation under depleted ammonium conditions (Luque et al., 2002; Sauer et al., 1999). We identified four iModulons, each overlapped with unique gene sets from the NtcA regulon (Takatani & Omata, 2006). We interpret these iModulons with consideration of their driving factors such as the different levels of nitrogen deficiency, the type of secondary nitrogen source available in aquatic environments, and the degree of photosynthetic demands.

The Nitrogen Assimilation iModulon captures 10 out of 28 genes known to be regulated by NtcA. These included the ABC-type nitrite and nitrate transporters (*nrtABCD*), nitrogen reductases (*nirA* and *narB*), cyanate transporter and its catabolic enzyme cyanase (*cynABDS*) (Espie et al., 2007). *S. elongatus* has the ability to assimilate and metabolize cyanate to saturate cellular pools of ammonium and CO_2 during nitrogen starvation (Maeda & Omata, 2009). We also observed the transcript of the ammonium transporter Amt1 below the iModulon threshold.

The second NtcA-regulated iModulon included two gene clusters, *Nit1C* and *cynS*, along with *nirA*, *nrtA* and 5 genes with unknown function. The gene cluster (*nitBCDEFGH*) in the canonical nitrilase Nit1C is highly conserved in nine different bacterial species (Jones et al., 2018; Podar et al., 2005). Studies have shown that Nit1C is essential for growth in cyanide (Estepa et al., 2012) while also being involved in nitrogen provision (Harris & Knowles, 1983; Jones et al., 2018), leading us to label this iModulon as Cyanide Assimilation (Adjei & Ohta, 1999; Buchanan et al., 2015; Finnegan et al., 1991; Harris & Knowles, 1983; Skowronski & Strobel, 1969; Estepa et al., 2012). However, the role of Nit1C has not been investigated in *S. elongatus*. Cyanate is formed by the oxidation of cyanide, but has different chemical properties and is metabolized in a separate pathway. Cyanase has not been shown to be associated with cyanide degradation to ammonium (Sáez et al., 2019). *nirA* was negatively weighted, suggesting that the expression of nitrite reductase is anti-correlated with the expression of putative cyanide metabolism. We also identified 5 additional uncharacterized genes in this iModulon that might be involved in cyanide metabolism or peptide synthesis (**Figure A.5c**).

The next iModulon describes the repressive role of NtcA applied to *gifA* and *gifB,* both encoding glutamine synthetase (GS) inactivating factors (García-Domínguez et al., 2000). Under nitrogen depletion, GS enzymatic activity is inhibited because of its dependency on ammonium to drive glutamine synthesis. This iModulon also captures *synpcc7942_1845*, encoding a hypothetical gene, that might also be involved in the NtcA-GS inhibition process.

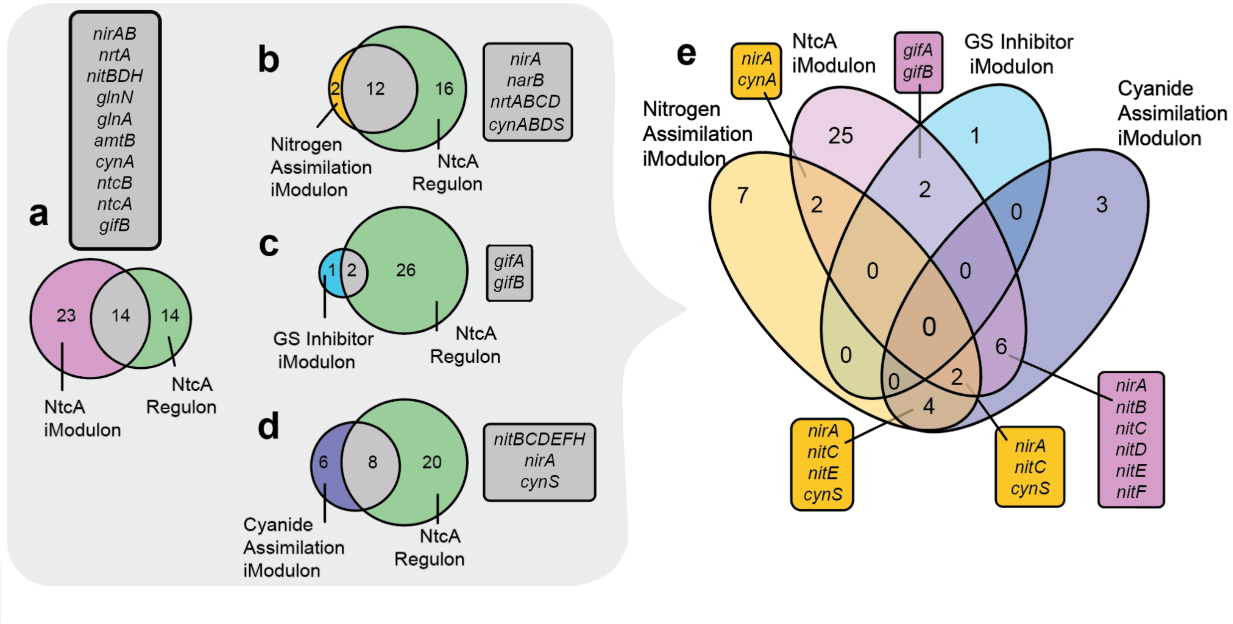


Figure 2.7 Different modes of regulation of the NtcA transcription factor. (a-d) iModulon Venn diagram with NtcA regulon. Each iModulon configuration and importance in described in the main text. The gray box next to every Venn diagram represents the shared genes between the indicated iModulon and the NtcA regulon **(e)** Four-way Venn diagram showing no overlap between the genes of each NtcA-regulated iModulons

The configuration of the last iModulon (labeled as NtcA) captures the global NtcA regulon with a list of conserved hypothetical genes. Neighboring genes to the captured hypothetical genes are involved in arginine decarboxylase, proteolysis, *a*midoligases, and amidotransferases (Karp et al., 2019). To understand the correlation of all four iModulon, we compared their activity levels across all conditions in the compendia. As expected, the GS inhibiting iModulon produced an opposite activity to the Nitrogen and Cyanide Assimilation iModulons. Interestingly, the repression of GS inhibiting iModulon is most active in the dark either because of deactivated photosynthesis (which is a nitrogen-dependent process), or the TCA cycle is deactivated TCA at night since gluconeogenesis is the main energy-producing metabolic pathway. Subjective dusk produced almost no expression across all the NtcA iModulons but that was not the case during subjective dawn, where positive activity between 5 to 30 mins was observed to possibly jump-start photosynthetic activity since most of its apparatus proteins depend on nitrogen.

2.5 Differential Activation of iModulons Biologically Classifies Uncharacterized iModulons

A significant region of *S. elongatus'* genome is uncharacterized, including 40% hypothetical genes and 15% unannotated essential genes (Rubin et al., 2015). We can inform the process of gene annotation through identifying differentially activated iModulons, or DIMA (Differential iModulon Activity), across two conditions of interest. We elucidate the utility of DIMA through a comparative analysis between two conditions with circadian times CT = 0 (subjective dawn, sample name: ppGpp:0wt_dusk_720) and CT = 12 (subjective dusk, sample name: ppGpp:0wt_dusk_000) to reveal a set of highly differentiated iModulons between both circadian timeframes (PRJNA:404081). From the DIMA plot for these conditions (**Figure 2.8a**), we were able to characterize two iModulons that were initially uncharacterized, labeled as uncharacterized-6 and uncharacterized-7 (or U5 and U7, respectively).

U5 contains a large sum of unannotated genes and growth-related genes, such as those encoding ribosomal subunits, biosynthetic enzymes and RNase II, to maintain a steady translation rate and cell viability in *S. elongatus* (Puszynska & O'Shea, 2017b)**.** This iModulon is highly expressed at dawn. Additionally, U7 contains replication and translation genes and other conserved hypothetical genes. This iModulon was downregulated at dawn (**Figure 2.8a**), which is opposite to the expression of U6. We hypothesize that U7 might be influence by ppGpp regulation so we named it as ppGpp-related iModulon. The content of both iModulons are useful tools for future studies to identify what genes in the genome are highly activated/repressed at these circadian times. These iModulon also include a large sum of poorly annotated genes, which can be further studied to fully characterize them since they might be related to the set of annotated genes in their respective iModulons.

Furthermore, we also conducted DIMA between WT and ΔrpaA strains which showed a consistency in the RpaA iModulon, since RpaA~P induces circadian gene expression for class I genes at the onset of dusk and throughout the night (**Figure 2.8b-c**). Between 30 minutes to 4 hours into darkness, circadian-regulated iModulons such as the Competence iModulons, Pili-related iModulon, and State Transition iModulon were upregulated in WT, since ΔrpaA cannot exert circadian influence on gene expression. This result confirms that the competence mechanism is either directly or indirectly (through RpaA sigma factors and other possible TFs) is regulated by the circadian clock. Contrary to previous results reporting a decrease in *kaiBC* expression by ~3.5-fold in ΔrpaA mutants (Markson et al., 2013; Takai et al., 2006), our DIMA plots showed an increase in the KaiBC iModulon (containing both *kaiBC*) expression in the ΔrpaA mutants from 30 minutes to 4 hours in darkness. The Phototaxis iModulon (containing the *tax1* operon for photoreceptor proteins) is not circadian regulated as indicative by the high activity in the ΔrpaA strain (Yang et al. 2018).

Another activity-based analysis is DIMCA (Differential iModulon Clustering Activity) which identifies sets of iModulons with high activity correlation. In other words, DIMCA attempts to cluster iModulons that are likely to be activated together under common underlying environmental stimuli or a global transcription factor that activates multiple downstream regulatory units. We have identified three clusters (Pearson R > 0.8 and silhouette score between 0-0.6), each is descriptive of different levels of grouped regulatory activity (**Figure 2.8**

d-f). The first cluster demonstrates the dependency of the photosynthetic apparatus to nitrogen, an important nutrient element to maintain photosynthetic activity in *S. elongatus*. A negative correlation between the Cyanide Metabolism iModulon (regulated by NtcA) to the photosystem iModulon, suggesting that nitrogen starvation negatively impacts the expression of photosynthetic apparatus (**Figure 2.8d**). Additionally, a similar correlation is observed between the Cyanide Metabolism iModulon and a single gene iModulon encoding a high-light inducible gene (*synpcc7942_01120)*. This might suggest that during high-light stress, *synpcc7942_01120* is expressed to mitigate photo-damage as part of S. elongatus' high-light acclimation process. While doing so, nitrogen supplementation to the photosynthetic apparatus is stopped, since *S. elongatus* will prioritize responding to the high-light stress over synthesizing chlorophyll pigments, which relies on sufficient nitrogen levels.

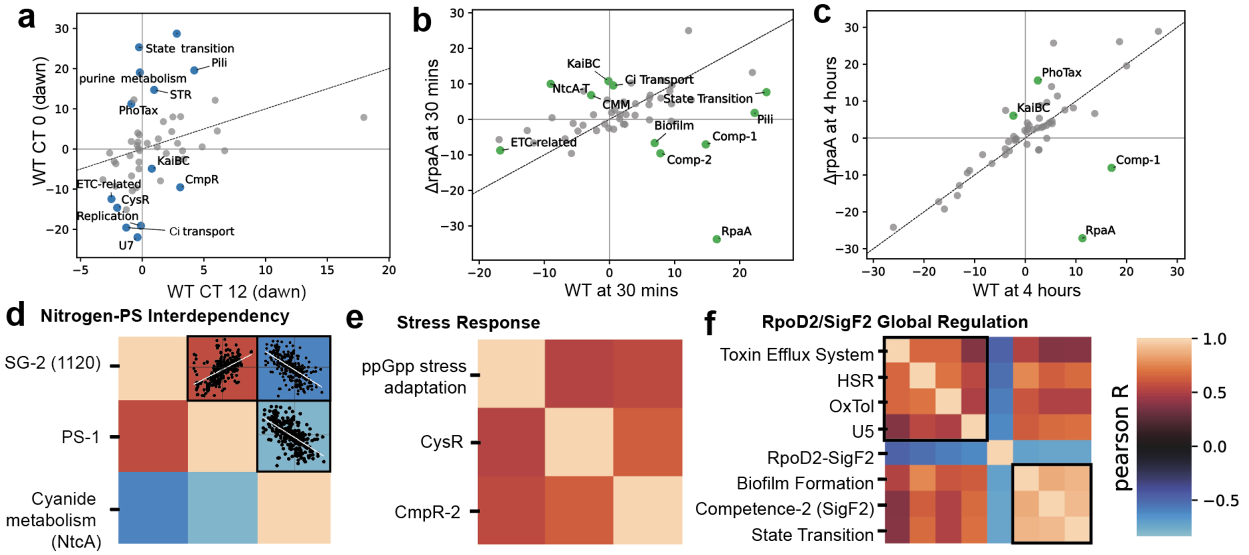


Figure 2.8: DIMA and DIMCA plots. (a) DIMA scatter plot comparison between dawn and dusk samples at CT 0 and CT 12, respectively. Every scatter point represents an iModulon activity across both indicated conditions. Samples were taken from PRJNA:404081. **(b-c)** DIMA scatter plot comparison between wildtype and rpaA null mutant at 30 minutes and 4 hours post-dawn. Samples were taken from PRJNA:354335. **(d-f)** DIMCA plots of iModulons with correlated activity levels (Pearson R 0.85) that illustrate iModulon association. Vertical sidebar to the right indicates the range of Pearson R score. Abbreviations: HSR - heat shock response, OxTol - oxidative stress tolerance, Competence-2 - competence iModulon regulated by SigF2.

The second cluster groups CysR and CmpR-2 (two nutrient stress iModulons) with the ppGpp stress iModulon (**Figure 2.8e)**. This cluster might suggest a role of ppGpp in regulating the expression of sulfur and inorganic carbon acquisition during nutrient starvation conditions.

The third cluster describes iModulons that are impacted by an increase in the expression of *rpoD2* and the derepression of *sigF2 (***Figure 2.8f)**. The rpoD2-sigF2 iModulon describes the influence of RpaA on the expression of both *rpoD2* and *sigF2*: when *sigF2* is expressed *rpoD2* is depressed by RpaA and vice versa (Fleming & O'Shea, 2018). With this cluster, we can establish all iModulons that are implicated by rpoD2-sigF2 directly. For example, the bottom right boxed iModulons are activated with abundant SigF2, most of which are iModulons activated under low light conditions such as natural competence, biofilm development, and state transition. The top left iModulons are those that are also expressed with abundant *sigF2* transcription. This also suggests that sigF2 or RpoD2 might have some regulatory influence on these iModulons. For example, the Competence-2 iModulon contains competence-participating genes such as *sigF2* and *drpA*, both of which have been previously associated with DNA uptake and protection (Taton et al. 2020). However, we were interested in the Toxin Efflux System iModulon that contains transcripts of chaperon proteins, secretion transporters, and SigG, whose protein function is uncharacterized in *S. elongatus*. We here purpose that *sigG* could be regulated by SigF directly. This iModulon also showed high correlation to the Oxidative Stress Tolerance (OxTol) iModulon and the Heat Shock Response (HSR) iModulon. This association has been previously described in an engineered *S. elongatus* strain with acyl-ACP synthetase removed and acyl-ACP thioesterase expressed for free-fatty acid secretion (Ruffing and Jones 2012). The engineered strain showed a high expression of chaperon proteins, SigG, secretion pumps, oxidoreductases and heat shock proteins. Therefore, DIMCA analysis is useful when studying correlated biological systems to elucidate their connections through transcriptional regulation.

2.6 iModulon analysis of time-course data confirms integrated regulatory system

Differential gene expression is commonly used to study the transcriptome. Here, we use iModulons to provide a structural basis to identify differentially activated collections of co-expressed genes (i.e. those that are collectively activated due to a common stimuli). We clustered significantly active iModulons with absolute values larger than 10 to investigate integrated regulatory networks that have similar activation patterns using time-course data, and thus inform orchestrated biological processes within *S. elongatus*. To achieve such a comparison, we generated two clustermaps: the first containing light conditions **(Figure 2.9)**, and the seconds between WT and rpaA null mutants **(Figure 2.10).** From these clustermaps, we confirmed that Comp-2 iModulon is regulated by the circadian clock through SigF2 because of how similar it behaves to the RpaA iModulon. Additionally, Uncharacterized iModulon 2 (U2) showed similar activity to the nitrogen assimilation, sulfur assimilation, and IdiB iModulons. From this, there is a possibility that U2 could respond to a particular nutrient stress. The second clustermap indicates that the RpaA iModulon behaves the closest to the photosystem iModulon, since photosynthesis is dependent of the circadian rhythm.

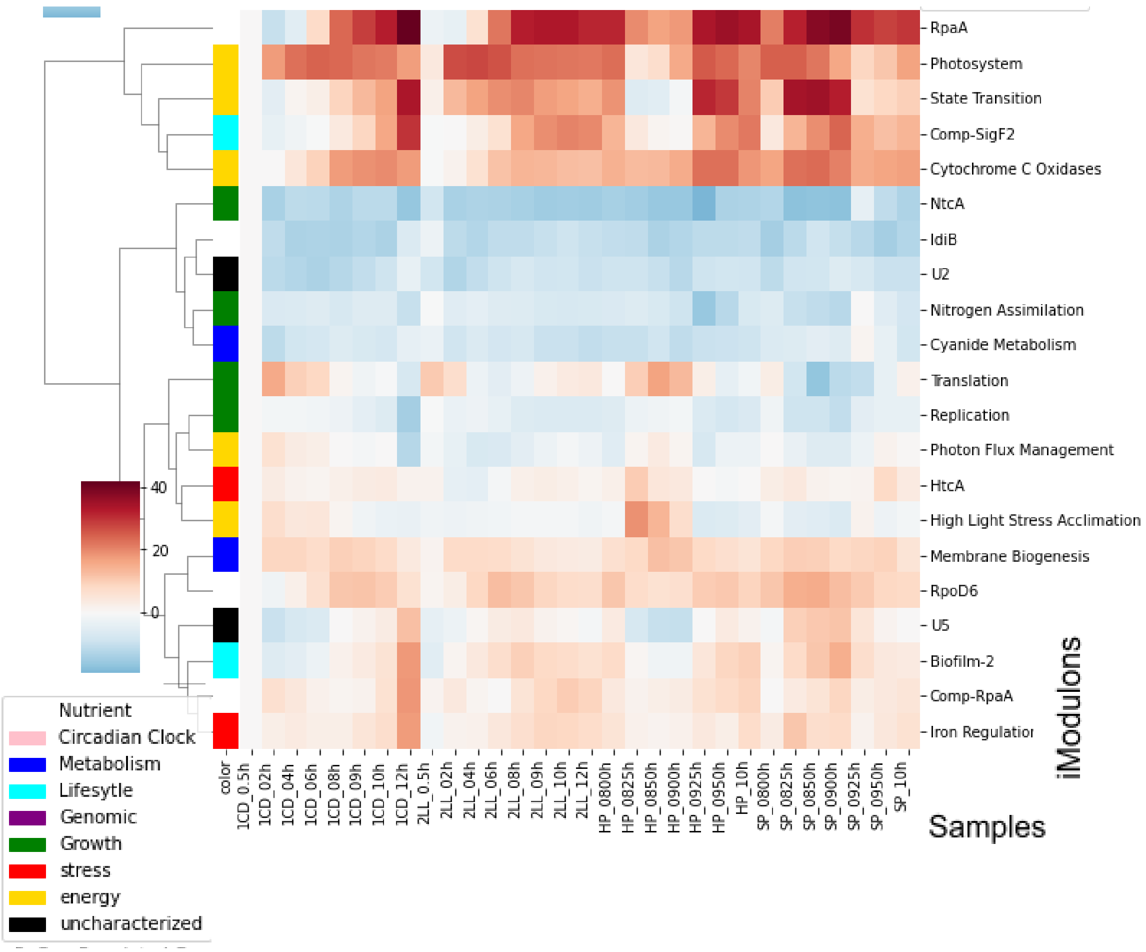


Figure 2.9: Clustermap of inferred iModulon activities across light conditions. Abbreviations: CD - clearday condition, HP - high light pulse condition, LL - low light condition, and SP - shade pulse condition. All samples were obtained from PRJNA:412032. iModulons are categorized based on their biological function and these categories are shown in the legend panel (bottom left).

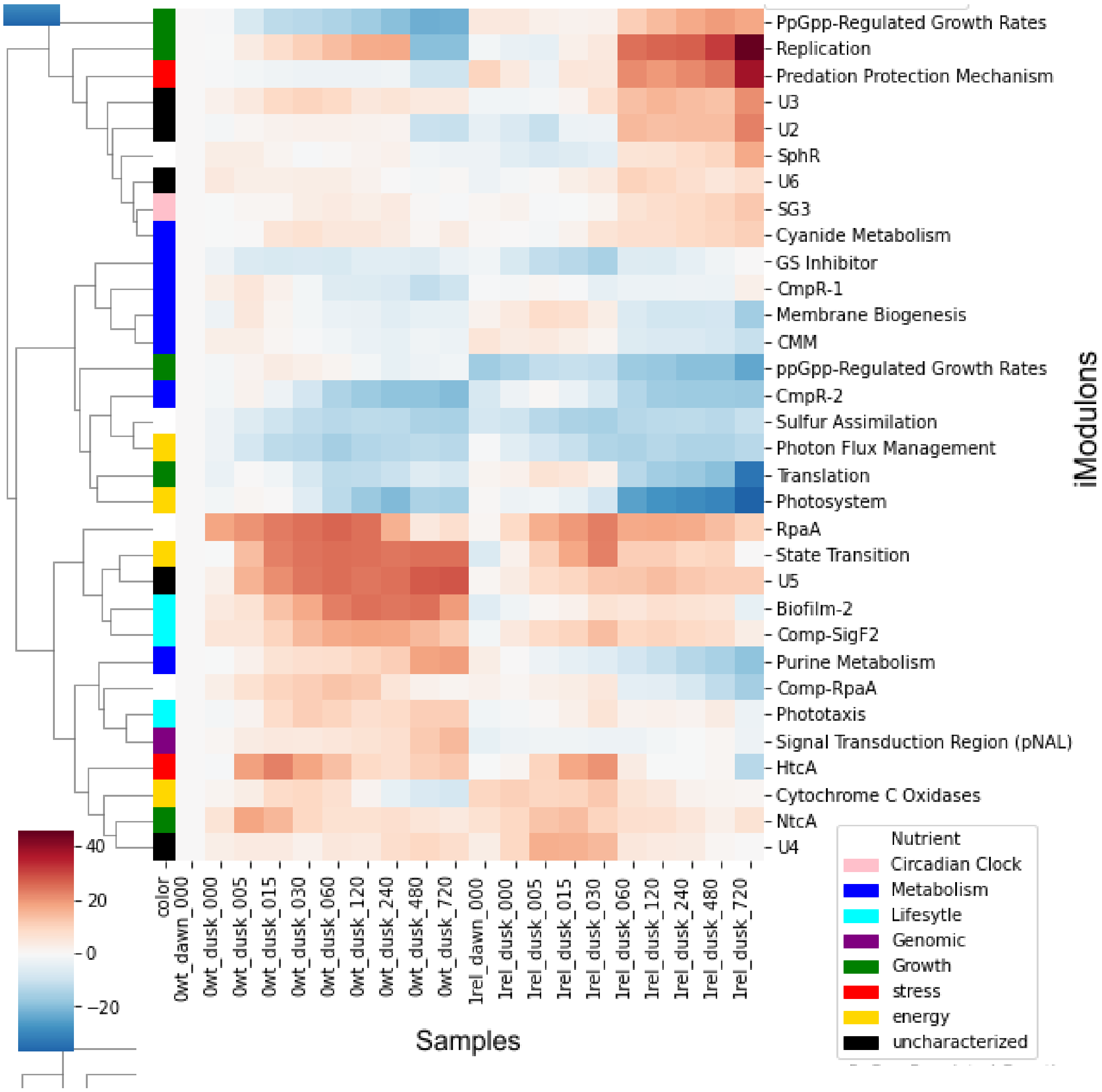


Figure 2.10: Clustermap of iModulon activities across WT and *rel* knockout mutants. Samples represented are obtained from PRJNA:401742 and PRJNA:404081. iModulons are categorized based on their biological function and these categories are shown in the legend panel (bottom left).

Chapter 2 is currently being prepared for submission for publication of the material. **Tahani Al Bulushi**, Anand V Sastry, Kevin Rychel, Saugat Poudel, Reo Yoo, Siddharth Chauhan, Yuan Yuan, Cigdem Sancar, Richard Szubin, Bernhard Ø. Palsson, Susan Golden. (2021). “Machine learning reveals the transcriptional regulatory network and circadian dynamics of the cyanobacteria *Synechococcus elongatus* PCC 7942”. The thesis author is the primary author.

2.7 References

1. Adjei, M. D., & Ohta, Y. (1999). Isolation and characterization of a cyanide-utilizing Burkholderia cepacia strain. *World Journal of Microbiology & Biotechnology*, *15*(6), 699–704.

2. Battchikova, N., Eisenhut, M., & Aro, E.-M. (2011). Cyanobacterial NDH-1 complexes: novel insights and remaining puzzles. *Biochimica et Biophysica Acta*, *1807*(8), 935–944.
3. Buchanan, B. B., Gruissem, W., & Jones, R. L. (2015). *Biochemistry and Molecular Biology of Plants*. John Wiley & Sons.

4. Billis, K., Billini, M., Tripp, H. J., Kyrpides, N. C., & Mavromatis, K. (2014). Comparative transcriptomics between Synechococcus PCC 7942 and Synechocystis PCC 6803 provide insights into mechanisms of stress acclimation. PloS One, 9 (10), e109738.

5. Cheng, D., & He, Q. (2020). Iron Deficiency in Cyanobacteria. In Q. Wang (Ed.), *Microbial Photosynthesis* (pp. 181–196). Springer Singapore.

6. Diamond, S., Jun, D., Rubin, B. E., & Golden, S. S. (2015). The circadian oscillator in Synechococcus elongatus controls metabolite partitioning during diurnal growth. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(15), E1916–E1925.
7. Emlyn-Jones, D., Ashby, M. K., & Mullineaux, C. W. (1999). A gene required for the regulation of photosynthetic light harvesting in the cyanobacterium Synechocystis 6803. *Molecular Microbiology*, *33*(5), 1050–1058.

8. Espie, G. S., Jalali, F., Tong, T., Zacal, N. J., & So, A. K.-C. (2007). Involvement of the cynABDS operon and the CO2-concentrating mechanism in the light-dependent transport and metabolism of cyanate by cyanobacteria. *Journal of Bacteriology*, *189*(3), 1013–1024.

9. Estepa, J., Luque-Almagro, V. M., Manso, I., Escribano, M. P., Martínez-Luque, M., Castillo, F., Moreno-Vivián, C., & Roldán, M. D. (2012). The nit1C gene cluster of Pseudomonas pseudoalcaligenes CECT5344 involved in assimilation of nitriles is essential for growth on cyanide. *Environmental Microbiology Reports*, *4*(3), 326–334.

10. Finazzi, G., Barbagallo, R. P., Bergo, E., Barbato, R., & Forti, G. (2001). Photoinhibition of Chlamydomonas reinhardtii in State 1 and State 2: damages to the photosynthetic apparatus under linear and cyclic electron flow. *The Journal of Biological Chemistry*, *276*(25), 22251–22257.
11. Finazzi, G., & Forti, G. (2004). Metabolic Flexibility of the Green Alga Chlamydomonas reinhardtii as Revealed by the Link between State Transitions and Cyclic Electron Flow. *Photosynthesis Research*, *82*(3), 327–338.

12. Finnegan, I., Toerien, S., Abbot, L., Smit, F., & Raubenheimer, H. G. (1991). Identification and characterisation of an Acinetobacter sp. capable of assimilation of a range of cyano-metal complexes, free cyanide ions and simple organic nitriles. *Applied Microbiology and Biotechnology*, *36*(1), 142–144.

13. Fleming, K. E., & O'Shea, E. K. (2018). An RpaA-Dependent Sigma Factor Cascade Sets the Timing of Circadian Transcriptional Rhythms in Synechococcus elongatus. *Cell Reports*, *25*(11), 2937–2945.e3.

14. Forchhammer, K., & Selim, K. A. (2020). Carbon/nitrogen homeostasis control in cyanobacteria. *FEMS Microbiology Reviews*, *44*(1), 33–53.

15. García-Domínguez, M., Reyes, J. C., & Florencio, F. J. (2000). NtcA represses transcription of gifA and gifB, genes that encode inhibitors of glutamine synthetase type I from Synechocystis sp. PCC 6803. *Molecular Microbiology*, *35*(5), 1192–1201.

16. Ghassemian, M., & Straus, N. A. (1996). Fur regulates the expression of iron-stress genes in the cyanobacterium Synechococcus sp. strain PCC 7942. *Microbiology*, *142 (Pt 6)*, 1469–1476.

17. Hanaoka, M., Takai, N., Hosokawa, N., Fujiwara, M., Akimoto, Y., Kobori, N., Iwasaki, H., Kondo, T., & Tanaka, K. (2012). RpaB, another response regulator operating circadian clock-dependent transcriptional regulation in Synechococcus elongatus PCC 7942. *The Journal of Biological Chemistry*, *287*(31), 26321–26327.

18. Harris, R., & Knowles, C. J. (1983). Isolation and growth of a Pseudomonas species that utilizes cyanide as a source of nitrogen. *Journal of General Microbiology*, *129*(4), 1005–1011.

19. Jones, L. B., Ghosh, P., Lee, J.-H., Chou, C.-N., & Kunz, D. A. (2018). Linkage of the Nit1C gene cluster to bacterial cyanide assimilation as a nitrogen source. *Microbiology*, *164*(7), 956–968.

20. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551.

21. Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093.

22. Kato, H., Kubo, T., Hayashi, M., Kobayashi, I., Yagasaki, T., Chibazakura, T., Watanabe, S., & Yoshikawa, H. (2011). Interactions between histidine kinase NblS and the response regulators RpaB and SrrA are involved in the bleaching process of the cyanobacterium Synechococcus elongatus PCC 7942. *Plant & Cell Physiology*, *52*(12), 2115–2122.

23. Kim, Y.-S., Kim, J.-J., Park, S.-I., Diamond, S., Boyd, J. S., Taton, A., Kim, I.-S., Golden, J. W., & Yoon, H.-S. (2018). Expression of OsTPX Gene Improves Cellular Redox Homeostasis and Photosynthesis Efficiency in Synechococcus elongatus PCC 7942. *Frontiers in Plant Science*, *9*, 1848.

24. Kranzler, C., Rudolf, M., Keren, N., & Schleiff, E. (2013). Chapter Three - Iron in Cyanobacteria. In F. Chauvat & C. Cassier-Chauvat (Eds.), *Advances in Botanical Research* (Vol. 65, pp. 57–105). Academic Press.

25. Lamoureux, C. R., Decker, K. T., Sastry, A. V., McConn, J. L., Gao, Y., & Palsson, B. O. (2021). PRECISE 2.0 - an expanded high-quality RNA-seq compendium for Escherichia coli K-12 reveals high-resolution transcriptional regulatory structure. In *bioRxiv* (p. 2021.04.08.439047). https://doi.org/10.1101/2021.04.08.439047

26. Laudenbach, D. E., & Grossman, A. R. (1991). Characterization and mutagenesis of

sulfur-regulated genes in a cyanobacterium: evidence for function in sulfate transport. *Journal of Bacteriology*, *173*(9), 2739–2750.

27. López-Redondo, M. L., Moronta, F., Salinas, P., Espinosa, J., Cantos, R., Dixon, R., Marina, A., & Contreras, A. (2010). Environmental control of phosphorylation pathways in a branched two-component system. *Molecular Microbiology*, *78*(2), 475–489.

28. Luque, I., Zabulon, G., Contreras, A., & Houmard, J. (2002). Convergence of two global transcriptional regulators on nitrogen induction of the stress-acclimation gene nblA in the cyanobacterium Synechococcus sp. PCC 7942. In *Molecular Microbiology* (Vol. 41, Issue 4, pp. 937–947). https://doi.org/10.1046/j.1365-2958.2001.02566.x

29. Maeda, S.-I., Badger, M. R., & Dean Price, G. (2002). Novel gene products associated with NdhD3/D4-containing NDH-1 complexes are involved in photosynthetic CO2 hydration in the cyanobacterium, Synechococcus sp. PCC7942. In *Molecular Microbiology* (Vol. 43, Issue 2, pp. 425–435). https://doi.org/10.1046/j.1365-2958.2002.02753.x

30. Maeda, S.-I., & Omata, T. (2009). Nitrite transport activity of the ABC-type cyanate transporter of the cyanobacterium Synechococcus elongatus. *Journal of Bacteriology*, *191*(10), 3265–3272.

31. Markson, J. S., Piechura, J. R., Puszynska, A. M., & O'Shea, E. K. (2013). Circadian control of global gene expression by the cyanobacterial master regulator RpaA. *Cell*, *155*(6), 1396–1408.

32. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O., & Sastry, A. V. (n.d.). *Optimal dimensionality selection for independent component analysis of transcriptomic data*. https://doi.org/10.1101/2021.05.26.445885

33. Moronta-Barrios, F., Espinosa, J., & Contreras, A. (2012). In vivo features of signal transduction by the essential response regulator RpaB from Synechococcus elongatus PCC 7942. *Microbiology*, *158*(Pt 5), 1229–1237.

34. Mullineaux, C. W., & Emlyn-Jones, D. (2005). State transitions: an example of acclimation to low-light stress. *Journal of Experimental Botany*, *56*(411), 389–393.

35. Nakamoto, H., & Kojima, K. (2017). Non-housekeeping, non-essential GroEL (chaperonin) has acquired novel structure and function beneficial under stress in cyanobacteria. *Physiologia Plantarum*, *161*(3), 296–310.

36. Nicholson, M. L., Gaasenbeek, M., & Laudenbach, D. E. (1995). Two enzymes together capable of cysteine biosynthesis are encoded on a cyanobacterial plasmid. *Molecular & General Genetics: MGG*, *247*(5), 623–632.

37. Nodop, A., Pietsch, D., Höcker, R., Becker, A., Pistorius, E. K., Forchhammer, K., & Michel, K.-P. (2008). Transcript profiling reveals new insights into the acclimation of the mesophilic fresh-water cyanobacterium Synechococcus elongatus PCC 7942 to iron starvation. *Plant Physiology*, *147*(2), 747–763.

38. Ohkawa, H. (1998). The use of mutants in the analysis of the CCM in cyanobacteria. *Canadian Journal of Botany. Journal Canadien de Botanique*, *76*, 1025–1034.

39. Omata, T., Gohta, S., Takahashi, Y., Harano, Y., & Maeda, S. (2001). Involvement of a CbbR homolog in low CO2-induced activation of the bicarbonate transporter operon in cyanobacteria. *Journal of Bacteriology*, *183*(6), 1891–1898.

40. Omata, T., Price, G. D., Badger, M. R., Okamura, M., Gohta, S., & Ogawa, T. (1999). Identification of an ATP-binding cassette transporter involved in bicarbonate uptake in the cyanobacterium Synechococcus sp. strain PCC 7942. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(23), 13571–13576.

41. Pan, L.-L., Onai, K., Uesaka, K., Ihara, K., Natsume, T., Takatani, N., Ishiura, M., & Omata, T. (2016). Transcriptional regulation of CmpR, the LysR family protein involved in CO2-responsive gene regulation in the cyanobacterium Synechococcus elongatus. *Biomedical Genetics and Genomics*, *1*(5). https://doi.org/10.15761/bgg.1000123

42. Pérez-Pérez, M. E., Esther Pérez-Pérez, M., Martín-Figueroa, E., & Florencio, F. J. (2009). Photosynthetic Regulation of the Cyanobacterium Synechocystis sp. PCC 6803 Thioredoxin System and Functional Analysis of TrxB (Trx x) and TrxQ (Trx y) Thioredoxins. In *Molecular Plant* (Vol. 2, Issue 2, pp. 270–283). https://doi.org/10.1093/mp/ssn070

43. Piechura, J. R., Amarnath, K., & O'Shea, E. K. (2017). Natural changes in light interact with circadian regulation at promoters to control gene expression in cyanobacteria. *eLife*, *6*. https://doi.org/10.7554/eLife.32032

44. Podar, M., Eads, J. R., & Richardson, T. H. (2005). Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. *BMC Evolutionary Biology*, *5*, 42.

45. Poudel, S., Tsunemoto, H., Seif, Y., Sastry, A. V., Szubin, R., Xu, S., Machado, H., Olson, C. A., Anand, A., Pogliano, J., Nizet, V., & Palsson, B. O. (2020). Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 17228–17239.

46. Price, G. D. (2011). Inorganic carbon transporters of the cyanobacterial CO2 concentrating mechanism. *Photosynthesis Research*, *109*(1-3), 47–57.

47. Puszynska, A. M., & O'Shea, E. K. (2017a). *Author response: Switching of metabolic programs in response to light availability is an essential function of the cyanobacterial circadian output pathway*. eLife Sciences Publications, Ltd. https://doi.org/10.7554/elife.23210.022

48. Puszynska, A. M., & O'Shea, E. K. (2017b). ppGpp Controls Global Gene Expression in Light and in Darkness in S. elongatus. *Cell Reports*, *21*(11), 3155–3165.

49. Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A., & Golden, S. S. (2015). The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(48), E6634–E6643.

50. Ruffing, A. M. (2013). RNA-Seq analysis and targeted mutagenesis for improved free fatty acid production in an engineered cyanobacterium. *Biotechnology for Biofuels*, *6*(1), 113.

51. Rychel, K., Sastry, A. V., & Palsson, B. O. (2020). Machine learning uncovers independen tly regulated modules in the Bacillus subtilis transcriptome. *Nature Communications*, *11*(1), 6338.

52. Sáez, L. P., Cabello, P., Ibáñez, M. I., Luque-Almagro, V. M., Roldán, M. D., & Moreno-Vivián, C. (2019). Cyanate Assimilation by the Alkaliphilic Cyanide-Degrading Bacterium Pseudomonas pseudoalcaligenes CECT5344: Mutational Analysis of the cyn Gene Cluster. *International Journal of Molecular Sciences*, *20*(12). https://doi.org/10.3390/ijms20123008

53. Saito, M., Watanabe, S., Nimura-Matsune, K., Yoshikawa, H., & Nakamoto, H. (2020). Regulation of the groESL1 transcription by the HrcA repressor and a novel transcription factor Orf7.5 in the cyanobacterium Synechococcus elongatus PCC7942. *The Journal of General and Applied Microbiology*, *66*(2), 85–92.

54. Sastry, A. V., Dillon, N., Poudel, S., Hefner, Y., Xu, S., Szubin, R., Feist, A., Nizet, V., & Palsson, B. (2020). Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility. In *bioRxiv* (p. 2020.05.04.077271). https://doi.org/10.1101/2020.05.04.077271

55. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A., & Palsson, B. O. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*, *10*(1), 5536.

56. Sauer, J., Gorl, M., & Forchhammer, K. (1999). Nitrogen starvation in synechococcus PCC 7942: involvement of glutamine synthetase and NtcA in phycobiliprotein degradation and survival. *Archives of Microbiology*, *172*(4), 247–255.

57. Seino, Y., Takahashi, T., & Hihara, Y. (2009). The response regulator RpaB binds to the upstream element of photosystem I genes to work for positive regulation under low-light conditions in Synechocystis sp. Strain PCC 6803. *Journal of Bacteriology*, *191*(5), 1581–1586.

58. Seki, A., Hanaoka, M., Akimoto, Y., Masuda, S., Iwasaki, H., & Tanaka, K. (2007). Induction of a group 2 sigma factor, RPOD3, by high light and the underlying mechanism in Synechococcus elongatus PCC 7942. *The Journal of Biological Chemistry*, *282*(51), 36887–36894.

59. Shibata, M., Katoh, H., Sonoda, M., Ohkawa, H., Shimoyama, M., Fukuzawa, H., Kaplan, A., & Ogawa, T. (2002). Genes essential to sodium-dependent bicarbonate transport in cyanobacteria: function and phylogenetic analysis. *The Journal of Biological Chemistry*, *277*(21), 18658–18664.

60. Skowronski, B., & Strobel, G. A. (1969). Cyanide resistance and cyanide utilization by a strain of Bacillus pumilus. *Canadian Journal of Microbiology*, *15*(1), 93–98.

61. Sztukowska, M., Bugno, M., Potempa, J., Travis, J., & Kurtz, D. M., Jr. (2002). Role of rubrerythrin in the oxidative stress response of Porphyromonas gingivalis. *Molecular Microbiology*, *44*(2), 479–488.

62. Takai, N., Nakajima, M., Oyama, T., Kito, R., Sugita, C., Sugita, M., Kondo, T., & Iwasaki, H. (2006). A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria. In *Proceedings of the National Academy of Sciences* (Vol. 103, Issue 32, pp. 12109–12114). https://doi.org/10.1073/pnas.0602955103

63. Takatani, N., & Omata, T. (2006). Effects of PII deficiency on expression of the genes involved in ammonium utilization in the cyanobacterium Synechocystis sp. Strain PCC 6803. *Plant & Cell Physiology*, *47*(6), 679–688.

64. Taton, A., Erikson, C., Yang, Y., Rubin, B. E., Rifkin, S. A., Golden, J. W., & Golden, S. S. (2020). The circadian clock and darkness control natural competence in cyanobacteria. Nature Communications, 11(1), 1688.

65. The Gene Ontology Consortium, & The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. In *Nucleic Acids Research* (Vol. 47, Issue D1, pp. D330–D338). https://doi.org/10.1093/nar/gky1055

66. Toledano, M. B., & Huang, B. (2016). Microbial 2-Cys Peroxiredoxins: Insights into Their Complex Physiological Roles. *Molecules and Cells*, *39*(1), 31–39.

67. Vijayan, V., Jain, I. H., & O'Shea, E. K. (2011). A high resolution map of a cyanobacterial transcriptome. *Genome Biology*, *12*(5), R47.

68. Welkie, D. G., Rubin, B. E., Diamond, S., Hood, R. D., Savage, D. F., & Golden, S. S. (2 019). A Hard Day's Night: Cyanobacteria in Diel Cycles. *Trends in Microbiology*, *27*(3), 231–242.

69. Woodger, F. J., Badger, M. R., & Price, G. D. (2003). Inorganic carbon limitation induces transcripts encoding components of the CO(2)-concentrating mechanism in Synechococcus sp. PCC7942 through a redox-independent pathway. *Plant Physiology*, *133*(4), 2069–2080.

70. Yang, Y., Lam, V., Adomako, M., Simkovsky, R., Jakob, A., Rockwell, N. C., Cohen, S. E., Taton, A., Wang, J., Lagarias, J. C., Wilde, A., Nobles, D. R., Brand, J. J., & Golden, S. S. (2018). Phototaxis in a wild isolate of the cyanobacterium Synechococcus elongatus. Proceedings of the National Academy of Sciences of the United States of America, 115(52), E12378–E12387.

71. Yousef, N., Pistorius, E. K., & Michel, K.-P. (2003). Comparative analysis of idiA and isiA transcription under iron starvation and oxidative stress in Synechococcus elongatus PCC 7942 wild-type and selected mutants. *Archives of Microbiology*, *180*(6), 471–483.

Chapter 3

Methods

3.1 Data acquisition and RNA-seq processing

The data used in this study consisted of 323 RNA-seq datasets collected from NCBI Sequence Read Archive (SRA) repository. The in-house developed pipeline applies Entrez Direct (Kans, 2020) to search and extract all publicly available RNA-seq datasets into a compiled metadata. The script for this pipeline can be found on Github (https://github.com/avsastry/modulome_workflow/tree/main/download_metadata). The obtained metadata file was then loaded into our standardized RNA-seq processing pipeline that uses NextFlow v20.01.0 (Tommaso et al., 2017) to ensure data reproducibility. This pipeline is also available on Github (https://github.com/avsastry/modulome-workflow).

The RNA-seq pipeline can be summarized in the following steps. The raw FASTQ files were downloaded from NCBI using fastq-dump (https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump). Next, Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used for read trimming which was then followed by the application of FastQC. Bowtie(Langmead et al., 2009) was used to align the fragmented read sequences to the reference genome for *S. elongatus*. RSEQC (Wang et al., 2012) was applied to find read direction before using featureCounts(Picardi, 2015) to generate read counts. A final dataset of 317 profiles was the output of this pipeline that was normalized in units of log-transformed Transcripts per Million (log-TPM).

3.2 Quality control Pipeline and MetaData curation

The log-transformed dataset was processed further using our quality control (QC) pipeline to remove poor quality expression profiles not suitable for subsequent analyses. The QC process adapts five criteria (**Figure A.1a**): (1) FastQC that discards samples with bad statistical scores for per base sequence quality, per sequence quality scores, per base n content and adapter content. (2) FeatureCounts removes samples with less than 5x10^5 reads mapped to coding sequences. (3) Hierarchical clustering to identify poor sample correlation between individual samples in the dataset due to using unconventional RNA-seq library preparation such as ribosome sequencing and 3' or 5' end sequencing. (4) a manual process to curate metadata for every experimental sample through literature studies. These included strain description, base media, nutrient sources, experimental treatments such as light conditions, and growth stages if disclosed. Each sample was assigned a project name (identified by a unique BioProject ID) followed by a condition name. The project and condition name were separated by a colon (Project_name: Condition_name). Biological and technical replicates had similar condition names. Time-course conditions included a suffix denoting respective time stamps right after the condition name. For example "rpaA: wt_dark_015" would represent a wildtype condition that belongs to the rpaA project with a sample time of 15 mins after dark exposure. Notably, the following BioProjects were combined into one single project, named ppGpp, since they belong to the same GEO accession SuperSeries: PRJNA415380, PRJNA412032, PRJNA404081, PRJNA401777, and PRJNA401742. The last QC step (5) discarded samples with poor replicate correlation (Pearson R < 0.9) and no replicate or metadata, such as the "S2-7" project. However, in the case of *S. elongatus,* an exception was made for the time-series dataset that failed step 5 (*rpaA*, *clock*, and *ppGpp* with GEO accession GSE103463). We rescued these samples with (Pearson R < 0.8) to study the temporal activity of iModulons. The final dataset was then normalized to the reference condition within each project (see Table 1 in Results chapter). Thus, all activities within a unique project are relative to a defined baseline condition when studying differential activities.

3.3 Computing Independent Component analysis to identify robust components

ICA decomposes a gene expression matrix (**X**) into the independent components iModulon matrix (**M**) and their project-specific condition activities (**A**), as shown in equation 1. A detailed description of ICA performance is mentioned in the Supplementary Methods.

$$X = MA \qquad (1)$$

To obtain the robust number of independent components (ICs), OptICA (McConn et al., n.d.) coupled with the scikit-learn (v0.23.2) implementation of FastICA was performed on the project-specific log-TPM RNA-seq compendium for *S. elongatus* with 100 iterations with random seeds and a convergences tolerance of 10^{-7}. The resulting independent components (ICs) were clustered using DBSCAN (Ester et al., 1996) to identify robust ICs, using an epsilon of 0.1 and a minimum cluster seed size of 50. To account for identical with opposite signs, the following distance metric was used for computing the distance matrix:

$$d_{x,y} = 1 - ||\rho_{x,y}|| \qquad (2)$$

where $\rho_{x,\ y}$ is the Pearson correlation between components a and b. The final robust ICs were defined as the centroids of the computed cluster (McConn et al.). Given that the number of dimensions can profoundly affect ICA results, we applied ICA to the transcriptomic data multiple times, from the number of dimensions (i.e. the dataset size) between 10 to 260 with a step increase rate of 10. The optimal dimensionality was identified by comparing the number of ICs with single genes to the number of ICs correlated with the ICs in the largest dimension (called “final components") (Pearson R > 0.7). The optimal dimension was defined as the number of dimensions where the number of non-single gene ICs was equal to the number of final components in that deconvolution (McConn et al.). The optimal dimension for this study was 120 (**Figure A.1c**).

3.4 Estimating iModulon Activity for External Datasets

Since profiles in the expression dataset are time-stamped, we inferred the total project-specific condition activities matrix (**A’**) that incorporated failed samples from QC to infer overall iModulon activity and compare that to the behavior obtained from **A**. To infer iModulon activities of all samples in the compendium, **A’** was calculated by inverting **M** and multiplying the resulting matrix to the log-TPM-norm or complete gene expression matrix (**X**) as shown below:

$$A' = M^{-1} X' \qquad (3)$$

3.5 Compiling the TRN and Gene Annotations

Regulon information for S. elongatus was manually curated from Biocyc (Karp et al., 2019) and from the literature that reported potential and ChIP-seq TF-DNA binding events. Transcription factor modes of effect (i.e. activation or repression) were reported in the TRN data. If marked as unknown, then the mode of effect was either not reported or its current mode of effect is poorly understood. With every regulator obtained, a regulon set, containing a list of all genes regulated by that specific regulator, was documented. Similarly, Gene annotations were pulled from AL009126.3. We also included a Cluster of Orthologous Groups (COG) and KEGG information using EggNOG mapper, Gene Ontology (GO) annotations using AmiGO2 (The Gene Ontology Consortium & The Gene Ontology Consortium, 2019), Uniprot IDs using the Uniprot ID mapper ("UniProt: The Universal Protein Knowledgebase in 2021," 2021), and operon clusters from Biocyc. The gene annotation pipeline can be found at (https://github.com/SBRG/pymodulon/blob/master/docs/tutorials/creating_the_gene_table.ipynb)

3.6 Computing iModulon Enrichments

iModulon enrichments against known regulons were computed using two-sided Fisher's exact test, with the FDR controlled at 10^{-5} using the Benjamini-Hochberg correction. Functional enrichment through KEGG and GO annotations were similarly computed but with FDR < 0.01. The supplementary packet includes all enrichments obtained in this study. By default, iModulons were compared to all possible single regulons and all possible combinations of two regulons to yield significant enrichments. If an iModulon yielded a significant KEGG or GO enrichment but no regulon enrichment, the iModulon was categorized as functional and was characterized via extensive literature studies. If none of these adjustments yielded a significant enrichment, the iModulon was annotated as non-regulatory (or either genomic, single gene, or uncharacterized).

iModulons annotated not using the default enrichment setup are noted in the iModulon table available as part of the included dataset.

3.7 Differential iModulon Activity Analysis

Differentially activated iModulons were computed across relevant conditions by using a log-normal probability distribution. iModulons with a difference greater than 5 and FDR < 0.01 were considered significant. DIMA 2D scatter plots compare the activity of all iModulons between two given conditions. Furthermore, DIMA for all-to-all condition comparisons can be clustered (DIMCA) used to identify globally regulated iModulons under a certain external stimulus, Stimulon. The scikit-learn agglomerative clustering function was implemented to create the cluster maps from the Seaborn package using equation 2 as the distance metric, where $\|\rho_{x,y}\|$ is the absolute value of the Pearson R correlation between two iModulon activity profiles. The threshold for optimal clustering was determined by testing different distance thresholds to locate the maximum silhouette score (> 0.6) (**Figure A.6**). Once established, the clusters were manually inspected to determine physiological function.

3.8 Identifying Homologous iModulons between Organisms

Considering a significant number of *S. elongatus*' genomes to be uncharacterized, we characterized a few iModulons through an iModulon comparison across different microbial species. Imodulon homology was identified by using equation 2 as the distance metric, where $\|\rho_{x,y}\|$ is the absolute value of the Pearson R correlation between two independent components. We considered two iModulons to be homologous or identical with a distance less than 0.25. Before comparing iModulons across species, gene orthology must be determined. To do this, we used reciprocal BLAST hits to generate one-to-one orthology between *S. elongatus* and other microbes. Once the orthologous pairs were determined, the iModulons were compared as described above.

3.9 Data and Code Availability

A list of all data used for this study and generated figures and tables are available at https://imodulondb.org. All code used to generate the results in this paper can be found on Github (https://github.com/talbulus/moduloome_selon). Custom code for our ICA analysis pipeline is also maintained on Github (https://github.com/SBRG/pymodulon).

Chapter 3 is currently being prepared for submission for publication of the material. **Tahani Al Bulushi**, Anand V Sastry, Kevin Rychel, Saugat Poudel, Reo Yoo, Siddharth Chauhan, Yuan Yuan, Cigdem Sancar, Richard Szubin, Bernhard Ø. Palsson, Susan Golden. (2021). “Machine learning reveals the transcriptional regulatory network and circadian dynamics of the cyanobacteria *Synechococcus elongatus* PCC 7942”. The thesis author is the primary author.

Chapter 3, in part, is material submitted for publication. Anand Sastry, Saugat Poudel, Kevin Rychel, Reo Yoo, Cameron Lamoureux, Siddharth Chauhan, Zachary B. Haiman, **Tahani Al Bulushi**, Yara Saif, and Bernard Ø. Palsson. (2021) Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. BioRxiv. DOI: https://doi.org/10.1101/2021.07.01.450581. The thesis author is the co-author.

3.10 References

1. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & Others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*, 226–231.

2. Kans, J. (2020). Entrez direct: E-utilities on the UNIX command line. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

3. Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093.

4. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

5. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O., & Sastry, A. V. (n.d.). *Optimal dimensionality selection for independent component analysis of transcriptomic data*. https://doi.org/10.1101/2021.05.26.445885

6. Picardi, E. (2015). *RNA Bioinformatics*. Springer New York.

7. The Gene Ontology Consortium, & The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. In *Nucleic Acids Research* (Vol. 47, Issue D1, pp. D330–D338). https://doi.org/10.1093/nar/gky1055

8. Tommaso, P. D., Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. In *Nature Biotechnology* (Vol. 35, Issue 4, pp. 316–319). https://doi.org/10.1038/nbt.3820
9. UniProt: The universal protein knowledgebase in 2021. (2021). *Nucleic Acids Research*, *49*(D1), D480–D489.

10. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* , *28*(16), 2184–2185.

Chapter 4

Discussion and Conclusion

Here, we used ICA to deconvolve 262 curated RNA-seq profiles of *S. elongatus* into 51 robust iModulon features, whose overall activity explained 73% of the variance in transcriptome across the wide variety of conditions contained in the dataset. 20 iModulons correspond to specific TFs, another 17 to specific biological functions, and the remaining 14 were either independent signals or lacked coherent biological interpretation. We analyzed the activity profiles associated with each enriched gene set in every iModulon and found that they either concur with existing knowledge or initiate data-driven hypotheses that could be experimentally validated in future studies. It is important to note that since *S. elongatus* lacks a properly defined regulon structure, iModulons serve as an unbiased, computational approach to reconstruct its TRN and identify new regulon structures in the process. Thus, this study presents the first and most thorough global TRN structure and corresponding activities for *S. elongatus*.

Through the application of ICA, we have demonstrated the efficiency of iModulon structures to overlap with well-defined regulons (including RpaA, CmpR-1, CmpR-2 and NtcA iModulons) and uncover biological insights that might add to the existing knowledge regarding the regulatory machinery in *S. elongatus.* We discovered a candidate Iron-related TF (*synpcc7942_2170*) that not only transcribes IdiB but also elucidates the regulatory crosstalk between IdiB and Fur. Moreover, autoregulated iModulons with helix-turn-helix domains were discovered through a comparison of *S. elongatus* to previously published iModulons to find new potential TFs and expand their regulons (CysR, HrcA, and Synpcc7942_0110)

Furthermore, we demonstrated how temporal data can be utilized to study the transcription of three photosynthesis-related iModulons under different light conditions. Temporal data was also used to cluster significantly activated iMOdulons (absolute expression of 10 or larger) to study iModulon activation across different timepoints.

ICA decomposition discovered multiple regulatory patterns for the global nitrogen assimilation by NtcA in response to photosynthetic demands and different nitrogen sources, including cyanide*.* We also demonstrated how iModulon analysis thought DIMA enables the characterization of, initially, poorly annotated iModulons.

Altogether, we have demonstrated that ICA extracts regulatory signals, in terms of iModulons, that correspond to either the mode-of-action underlying a TF or a sigma factor. These regulatory units can distinguish between biological machinery that is regulated by the circadian clock from those that are stimulated from environmental changes or metabolic shifts, which is an important distinction to make to optimize the metabolic and strain engineering of *S. elongatus.* ICA has also defined, for the first time, a complete TRN structure for *S. elongatus* from using only 317 RNA-seq databases which proves the importance of sequencing data-mining for biological discoveries at the transcription level and specifically, gene annotation through the co-regulated genetic features found in iModulons. However, ICA is not only limited to identifying differentially activated iModulons (DIMA), but also extends our understanding of system-level and global regulatory changes that influences a set of collective iModulons. This is known as Differential iModulon Cluster Activity (DIMCA) through which the identification of co-regulated iModulons induced by environmental stimulii, also known as “stimulons”, is made possible (Lamoureux et al., 2021). The evolution of multi-scale analytical tools (from DEGs to iModulons to stimulons) is extremely useful for the development of meticulous biochemical assays, where individual stimulons can be closely studied at the molecular level. Therefore, this study motivates the generation of additional *S. elongatus* RNA-seq data under a diverse collection of unique experimental conditions to enable a more precise rendering of its TRN, since most machine learning performance is limited to the diversity within the dataset (McConn et al.). Nevertheless, with the current datasize, the iModulons presented show high biological significance because they capture well-defined regulons and cellular pathways. iModulons also depict interdependent regulatory modules, hierarchical regulation, and potential regulon structures. In theory, if a complete transcriptomic dataset that reflects every considerable condition for *S. elongatus* were to be gathered, ICA would reveal a comprehensive TRN from its simplest building blocks to the highest levels of global regulation.

In conclusion, we introduced the utility of ICA as a tool to decompose gene expression data into modulated regulatory units that describe the response mechanism and dynamic behavior of *S. elongatus.* We used the largest publicly-available RNA-seq compendium and extensively annotated every condition to remain a reliable resource in the community. The 51 iModulons extracted from ICA validated existing knowledge while also revealing putative regulation, including transcription factor discovery and genotype-phenotype- relationships.

Chapter 2 presents the results of this study where we explored the most significant iModulons based on their composition and activity. We demonstrated the effectiveness of iModulons in simplifying the study of transcriptional changes, such as those induced via genetic perturbations and changes in experimental treatments and environmental dynamics. Changes in iModulon activity levels exhibited the capacity and potential of how ICA can interpret the changes in expression levels within the transcriptomic dataset. From the activity matrix, we were able to demonstrate how studying differential iModulon Activity (DIMA) is a simplified approach in studying differential gene expression. This knowledge tool serves as a resource before designing biochemical assays in the future. We encourage the community to refute to iModulon database to discover more of these iModulon structures, emphasizing the great benefits of big data analytics.

However, although ICA being a blind source separation is limited to only capturing linear regulatory interactions (excluding hierarchical and non-linear interactions) the analyses presented in this thesis demonstrate that ICA extracts accurate and interpretable information that describes *S. elongatus'* transcriptome. It is still ambiguous if the capacity of ICA is extendable to eukaryotic studies, considering their complex structures, but recent studies have shown promising results (Nazarov et al., 2018; Soheili-Nezhad et al., 2021; Wang et al., 2021). ICA can also be applied to other types of omic-data including proteomics.

The field of systems biology and big data is revolutionizing our approach to studying transcriptomics. They provide a mechanistic view of biological systems and are a promising platform to derive biological insights. Through this work, we have demonstrated that ICA enables the interpretation of big expression datasets by revealing robust biological signals that are propagated in the cells to induce gene expression. These signals can be studied further to better understand cellular dynamics and their response mechanism to these signals.

Chapter 4 is currently being prepared for submission for publication of the material. **Tahani Al Bulushi**, Anand V Sastry, Kevin Rychel, Saugat Poudel, Reo Yoo, Siddharth Chauhan, Yuan Yuan, Cigdem Sancar, Richard Szubin, Bernhard Ø. Palsson, Susan Golden. (2021). “Machine learning reveals the transcriptional regulatory network and circadian dynamics of the cyanobacteria *Synechococcus elongatus* PCC 7942”. The thesis author is the primary author.

5.1 References

1. Lamoureux, C. R., Decker, K. T., Sastry, A. V., McConn, J. L., Gao, Y., & Palsson, B. O. (2021). PRECISE 2.0 - an expanded high-quality RNA-seq compendium for Escherichia coli K-12 reveals high-resolution transcriptional regulatory structure. In *bioRxiv* (p. 2021.04.08.439047). https://doi.org/10.1101/2021.04.08.439047

2. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O., & Sastry, A. V. (n.d.). *Optimal dimensionality selection for independent component analysis of transcriptomic data*. https://doi.org/10.1101/2021.05.26.445885

3. Nazarov, P. V., Wienecke-Baldacchino, A. K., Zinovyev, A., Czerwińska, U., Muller, A., Nashan, D., Dittmar, G., Azuaje, F., & Kreis, S. (2018). Independent component analysis provides clinically relevant insights into the biology of melanoma patients. In *bioRxiv* (p. 395145). https://doi.org/10.1101/395145

Appendix A

Supplementary Information and Figures

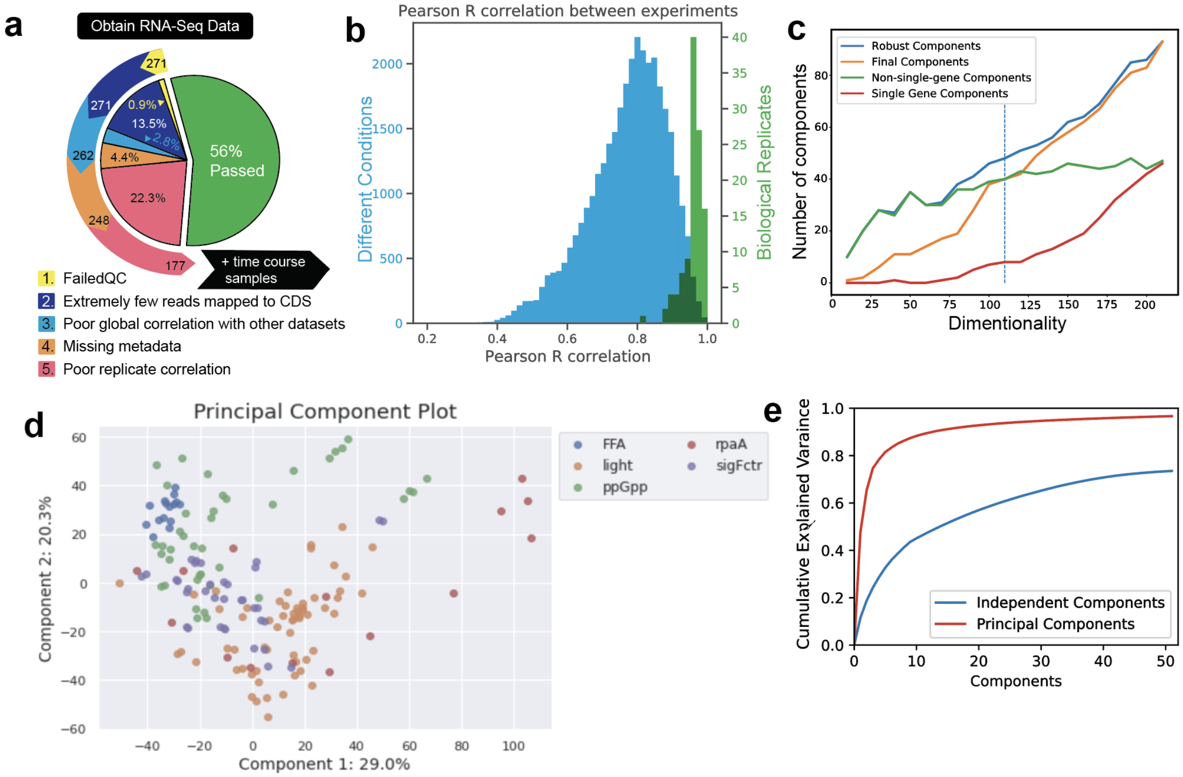


Figure A.1: Overview of *S. elongatus* dataset. (a) Pie chart showing quality control criterias and the number of RNA-seq data that passed each criteria. The number of passed and failed samples that represent 56% and 44% of the data are 177 and 140, respectively. This process was followed by manually adding all failed time-course dataset to increase the final passed sample size from 177 to 262. **(b)** Histogram of data quality, as measured by the coefficient of determination (R^2) between log-TPM. Comparisons between biological replicates are shown in green, whereas comparisons between all pairwise non-replicates are shown in blue. **(c)** Plot showing the number of components obtained from multiple dimensions (i.e. the dataset size) between 10 to 260 with a step increase rate of 10. The optimal dimensionality was identified by comparing the number of ICs with single genes to the number of ICs correlated with the ICs in the largest dimension (called “final components”)(Pearson R > 0.7). The optimal dimension, being 120 for this study, was defined as the number of dimensions where the number of non-single gene ICs was equal to the number of final components in that deconvolution.**(d)** Principal component analysis (PCA) plot showing the loadings of the first two principal components (PCs) of the passed 177 data after QC, colored by their respective project title (see Table 1 in the result section for more details). **(e)** Cumulative explained variance for the first 51 components calculated by principal component analysis (orange) and ICA (blue).

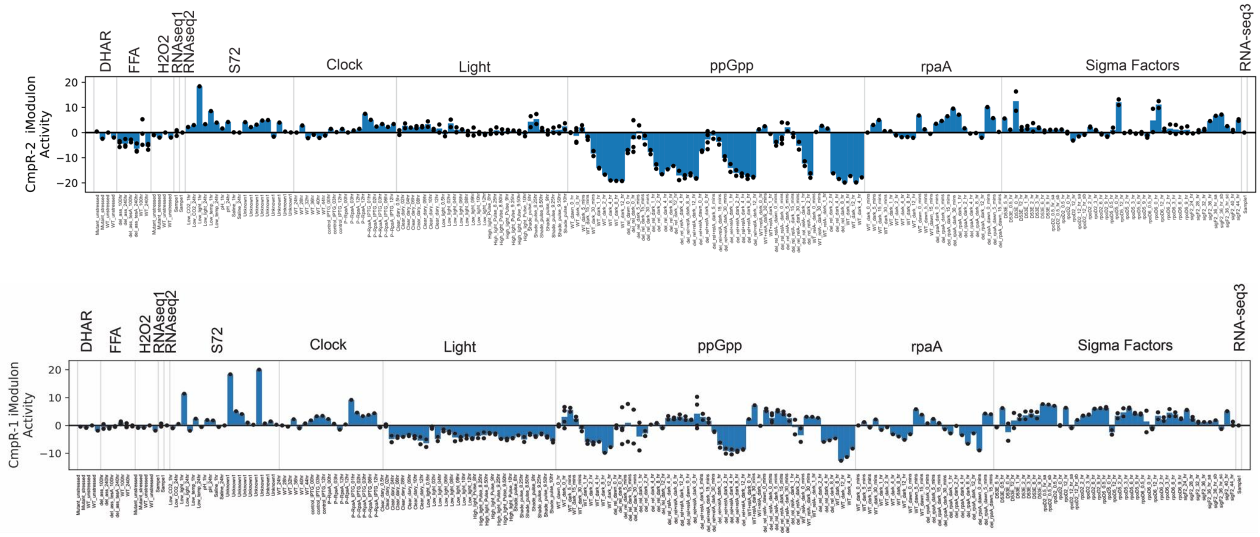


Figure A.2: CmpR-1 and CmpR-2 iModulon activity across all samples in the transcriptomic compendium. Project names are indicated above the activity bar (Table 1 contains more details regarding the original study for each project). Sample names are shown along the x-axis and the iModulon name is shown along the y-axis. For a full description of these activity plot, please visit iModulonDB.org.

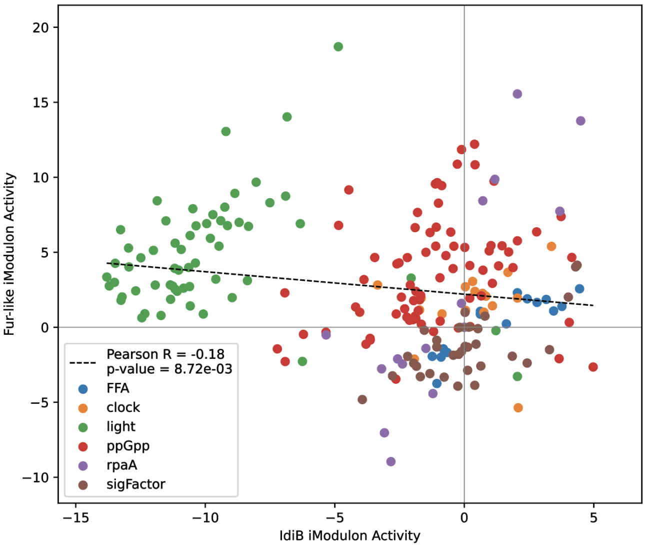


Figure A.3: Iron-related and IdiB iModulon activity comparison across all conditions. Highest variation occurs in the light project where IdiB iModulon shows negative activity and the Iron-related iModulon shows a positive upregulation. Further analysis on the regulatory cross-talk and the mode of effect of the proposed gene (*synpcc7942_2170*) encoding the regulatory transcription factor for the iron-related iModulon should be further investigated.

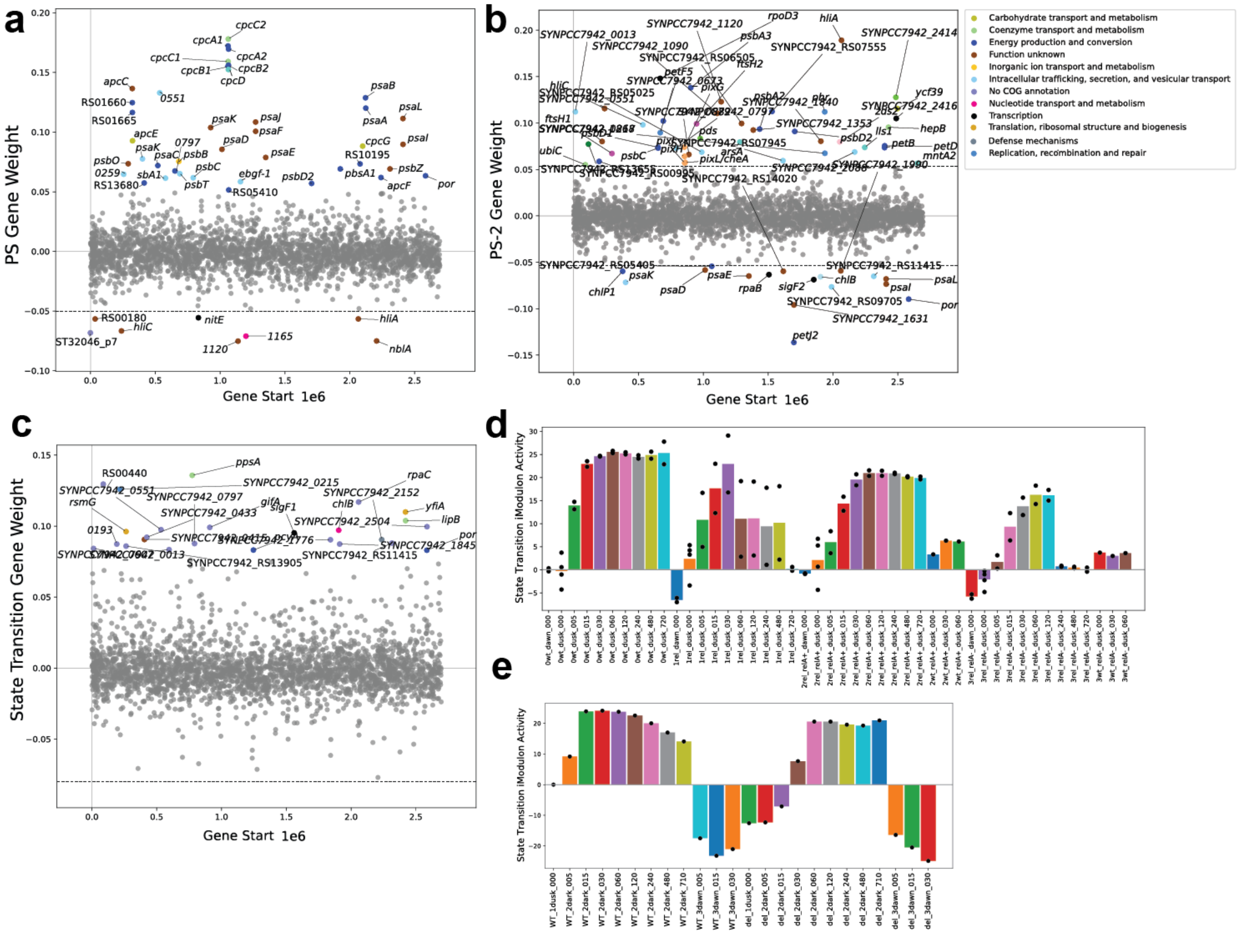


Figure A.4: RpaB-regulated iModulon gene weights and activity bar plots. **(a)** iModulon gene weights for PS iModulon including phycobilisomes, PSI and PSI reaction centers genetic features. **(b)** iModulon gene weight plot for the high-light stress acclimation (HLSA) iModulon. **(c)** iModulon gene weight plot for State Transition iModulon. **(d)** State Transition bar plots showing activity at night time for the ppGpp (PRJNA401742, PRJNA404081, PRJNA403840, PRJNA415380) project and **(e)** rpaA (PRJNA354335) Project.

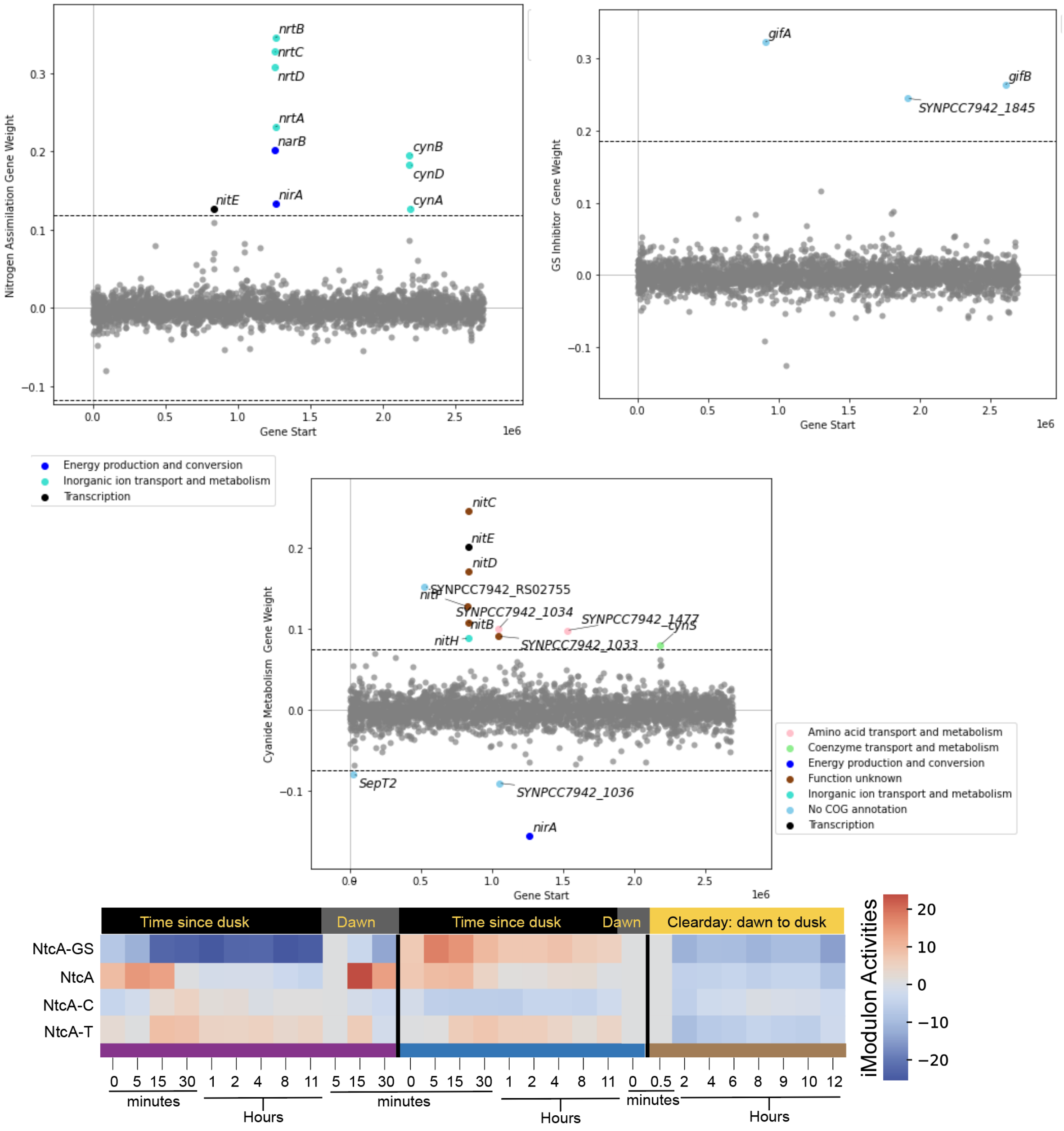
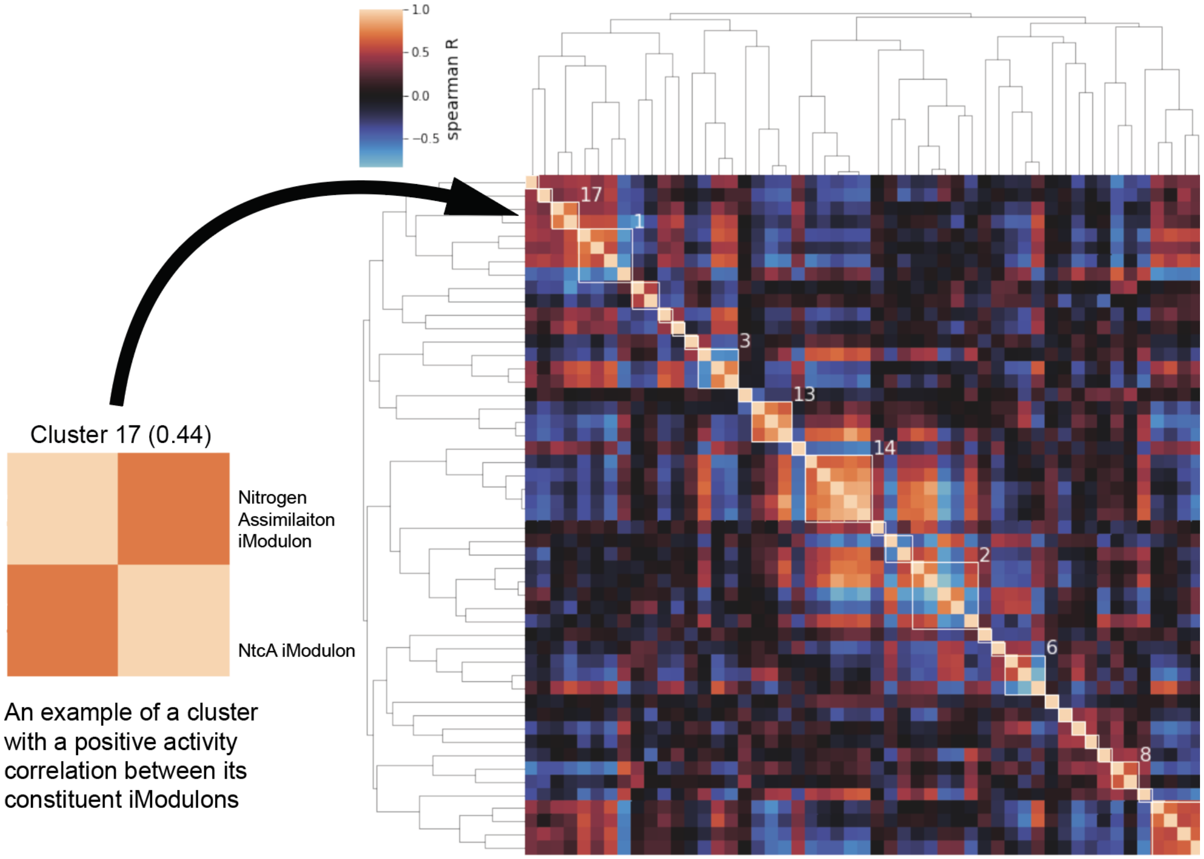


Figure A.5: NtcA- related iModulon configurations. (a) Nitrogen Assimilation iModulon **(b)** GS inhibitor iModulon regulated by NtcA, and © Cyanide Metabolism iModulon. **(d)** Clustermap comparing activities for all NtcA-related iModulons across different samples from the transcriptomic compendium.



An example of a cluster with a positive activity correlation between its constituent iModulons

Global clustering of all iModulons based on their activity with Pearson R cuttout of 0.5

Figure A.6: Clustering of iModulon activities using Pearson R to define correlated iModulons. Left cluster is the global clustermap of all iModulons in the transcriptomic compendium and right cluster shows an example of a “best cluster” with Pearson R correlation larger than the specified cutoff threshold (0.5 for this clustermap). Silhouette score for the right cluster is 0.44.

RpaA iModulon

Regulated by RpaA

Biological function: Circadian gene regulation

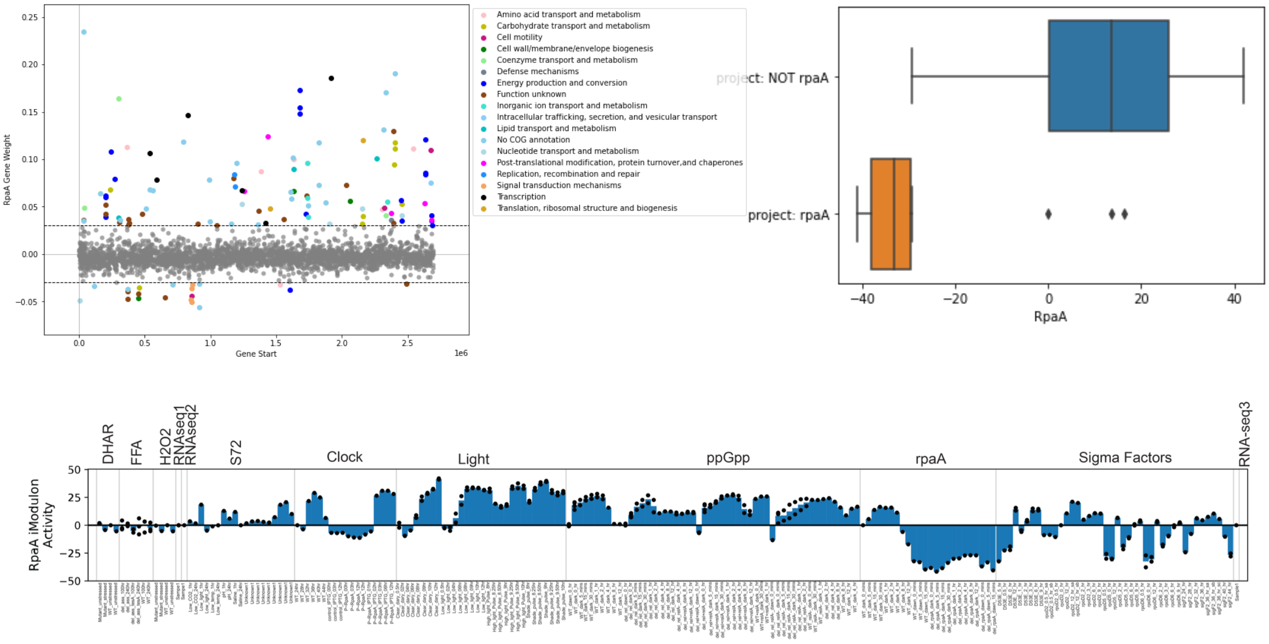


Figure A.7: Descriptive characteristics of the RpaA iModulon. (a) Scatterplot of iModulon gene coefficients against gene expression (log-TPM) in the reference condition. Genes are colored by their cluster of orthologous groups (COG) categories. **(b)** boxplot showing the different between mean expression of rpaA biopoject verse the remaining samples in the expression compendium. **(c)** iModulon activities across the entire compendium, grouped by original study as indicated above each project block (see Table 1 in the results section for more details). Each condition occupies a constant width, regardless of the number of biological replicates. All reference conditions are included in Table 1 in the Results section.

Note: Similar plots are shown for the remaining iModulons in the iModulon database: https://imodulondb.org