UNIVERSITY OF CALIFORNIA SAN DIEGO

The structure, dynamics and evolution of transcriptional regulation in *Staphylococcus aureus*

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

in

Biomedical Sciences

by

Saugat Poudel

Committee in charge:

Professor Bernhard Ø. Palsson, Co-Chair
Professor Victor Nizet, Co-Chair
Professor George Liu
Professor Joseph Pogliano
Professor Dong Wang
Professor Karsten Zengler

2022

The dissertation of Saugat Poudel is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my mom Sandhya and my wife Jamie, the two lights of my life

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors Professors Bernhard Ø. Palsson and Victor Nizet. It is rare to have the privilege of working with two fantastic advisors each providing perspective from basic biology to clinical importance. Much of the success in my graduate career can be attributed to the culture of open dialog, collaboration and creative scientific exploration they have established in their respective labs. I would also like to thank all of my committee members, whose inputs have been invaluable in bringing this dissertation to fruition. This work also would not have been possible without the support of Richard Szubin and Ying Hefner who have generated much of the experimental data that this research is built on.

In my time in graduate school I have also been fortunate enough to have several peer-mentors throughout. I would like to specially thank Nina Gao, Nick Dillon, Anand Sastry, Erol Kavvas, Yara Seif, Hannah Tsunemoto and Jon Monk for their close guidance and friendship. I am also grateful to have been part of a special cohort in SBRG and the Biomedical Sciences program, where I have found lifelong friendships. I want to thank CJ Norsigian, Aaron Oom, Michael Valdez, Ryan Geusz, Emily Griffin, Cedric Snethlage and Chris Park for sharing in this long journey through graduate school.

I am also grateful to all the innumerable people from throughout my life who have made it possible for me to complete this lifelong dream. I would like to thank Lance Winmill and Keith Eager, who believed in me from early on and instilled in me the courage to pursue my dreams. A special thank you to my undergraduate mentor Professor Wendy E. Thomas who took a chance on me and paved the way to my far fetched dream of becoming a scientist. Lastly, I would like to thank my brother Sumiran, my mom Sandhya and my wife Jamie for their unconditional love and support throughout.

Chapter 2, in part, is a reprint of material published in: **Poudel S**, Tsunemoto H, Seif Y, Sastry AV, Szubin R, Xu S, Machado H, Olson C, Anand A, Pogliano J, Nizet V, Palsson B Ø. Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. Proc Natl Acad Sci U S A. 2020;117: 17228–17239. The dissertation author was the primary author.

Chapter 3, in part, is currently being prepared for submission for publication: **Poudel S**, Hefner Y, Szubin R, Sastry A, Gao Y, Nizet V, and Palsson B Ø. "Coupling of CcpA and CodY activities coordinates carbon and nitrogen metabolism associated gene expression in *S. aureus* USA300 strains." The dissertation author was the primary author.

Chapter 4, in part, is currently being prepared for submission for publication: **Poudel, S**, Hyun J, Hefner Y, Nizet V, Palsson B Ø. "Interpreting roles of mutations in the emergence of S. aureus USA300 strains with genetics and independent component analysis of gene expression." The dissertation author was the primary author.

VITA

2015	Bachelor of Science in Microbiology, University of Washington
2022	Doctor of Philosophy in Biomedical Sciences, University of California San Diego

PUBLICATIONS

Kavvas ES, Seif Y, Yurkovich JT, Norsigian C, **Poudel S**, Greenwald WW, et al. Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions. BMC Syst Biol. 2018;12: 25.

Seif Y, Monk JM, Mih N, Tsunemoto H, **Poudel S**, Zuniga C, et al. A computational knowledge-base elucidates the response of Staphylococcus aureus to different media types. PLoS Comput Biol. 2019;15: e1006644.

Poudel S, Tsunemoto H, Meehan M, Szubin R, Olson CA, Lamsa A, et al. Characterization of CA-MRSA TCH1516 exposed to nafcillin in bacteriological and physiological media. Sci Data. 2019;6: 43.

Gao NJ, Al-Bassam MM, **Poudel S**, Wozniak JM, Gonzalez DJ, Olson J, et al. Functional and Proteomic Analysis of Streptococcus pyogenes Virulence Upon Loss of Its Native Cas9 Nuclease. Front Microbiol. 2019;10: 1967.

Anand A, Chen K, Yang L, Sastry AV, Olson CA, **Poudel S**, et al. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. Proc Natl Acad Sci U S A. 2019;116: 25287–25292.

Rajput A, **Poudel S**, Tsunemoto H, Meehan M, Szubin R, Olson CA, et al. Profiling the effect of nafcillin on HA-MRSA D712 using bacteriological and physiological media. Sci Data. 2019;6: 322.

Seif Y, **Poudel S**, Tsunemoto H, Szubin R, Meehan MJ, Olson C, et al. Profiling the effect of nafcillin on HA-MRSA D592 using bacteriological and physiological media. bioRxiv. 2020. p. 2020.04.30.070904. doi:10.1101/2020.04.30.070904

Rajput A, **Poudel S**, Tsunemoto H, Meehan M, Szubin R, Olson CA, et al. Identifying the effect of vancomycin on HA-MRSA strains using bacteriological and physiological media. bioRxiv. 2020. p. 2020.05.06.079640. doi:10.1101/2020.05.06.079640

Dillon NA, Seif Y, Tsunemoto H, **Poudel S**, Meehan M, Szubin R, et al. Characterizing the response of Acinetobacter baumannii ATCC 17978 to azithromycin in multiple in vitro growth conditions. bioRxiv. 2020. p. 2020.05.19.079962. doi:10.1101/2020.05.19.079962

Poudel S, Tsunemoto H, Seif Y, Sastry AV, Szubin R, Xu S, et al. Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. Proc Natl Acad Sci U S A. 2020;117: 17228–17239.

Sastry AV, Hu A, Heckmann D, **Poudel S**, Kavvas E, Palsson B Ø. Independent component analysis recovers consistent regulatory signals from disparate datasets. PLOS Computational Biology. 2021. p. e1008647. doi:10.1371/journal.pcbi.1008647

Rychel K, Decker K, Sastry AV, Phaneuf PV, **Poudel S**, Palsson B Ø. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. Nucleic Acids Res. 2021;49: D112–D120.

Gao Y, **Poudel S**, Seif Y, Shen Z, Palsson B Ø. Elucidating the CodY regulon in Staphylococcus aureus USA300 substrains. bioRxiv. 2021. p. 2021.01.08.426013. doi:10.1101/2021.01.08.426013

Rajput A, **Poudel S**, Tsunemoto H, Meehan M, Szubin R, Olson CA, et al. Identifying the effect of vancomycin on health care–associated methicillin-resistant Staphylococcus aureus strains using bacteriological and physiological media. Gigascience. 2021;10. doi:10.1093/gigascience/giaa156

Rajput A, Seif Y, Choudhary KS, Dalldorf C, **Poudel S**, Monk JM, et al. Pangenome Analytics Reveal Two-Component Systems as Conserved Targets in ESKAPEE Pathogens. mSystems. 2021;6. doi:10.1128/mSystems.00981-20

Sastry AV, **Poudel S**, Rychel K, Yoo R, Lamoureux CR, Chauhan S, et al. Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks. bioRxiv. 2021. p. 2021.07.01.450581. doi:10.1101/2021.07.01.450581

Yoo R, Rychel K, **Poudel S**, Al-bulushi T, Yuan Y, Chauhan S, et al. Machine learning of all Mycobacterium tuberculosis H37Rv RNA-seq data reveals a structured interplay between metabolism, stress response, and infection. bioRxiv. 2021. p. 2021.07.01.450045. doi:10.1101/2021.07.01.450045

Sastry AV, Dillon N, Anand A, **Poudel S**, Hefner Y, Xu S, et al. Machine Learning of Bacterial Transcriptomes Reveals Responses Underlying Differential Antibiotic Susceptibility. mSphere. 2021;6: e0044321.

Choe D, Szubin R, **Poudel S**, Sastry A, Song Y, Lee Y, et al. RiboRid: A low cost, advanced, and ultra-efficient method to remove ribosomal RNA for bacterial transcriptomics. PLoS Genet. 2021;17: e1009821.

Chauhan SM, **Poudel S**, Rychel K, Lamoureux C, Yoo R, Al Bulushi T, et al. Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile Sulfolobus acidocaldarius. Front Microbiol. 2021;12: 753521.

McConn JL, Lamoureux CR, **Poudel S**, Palsson B Ø, Sastry AV. Optimal dimensionality selection for independent component analysis of transcriptomic data. BMC Bioinformatics. 2021;22: 584.

Yuan Y, Seif Y, Rychel K, Yoo R, Chauhan S, **Poudel S**, et al. Pan-genomic analysis of transcriptional modules across Salmonella Typhimurium reveals the regulatory landscape of different strains. bioRxiv. 2022. p. 2022.01.11.475931. doi:10.1101/2022.01.11.475931

Xavier JB, Monk JM, **Poudel S**, Norsigian CJ, Sastry AV, Liao C, et al. Mathematical models to study the biology of pathogens and the infectious diseases they cause. iScience. 2022;25: 104079.

ABSTRACT OF THE DISSERTATION

The structure, dynamics and evolution of transcriptional regulation in *Staphylococcus aureus*

by

Saugat Poudel

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2022

Professor Bernhard Ø. Palsson, Co-Chair
Professor Victor Nizet, Co-Chair

Staphylococcus aureus is a versatile pathogen and a leading urgent threat to human health. The clinical burden of this organism is predicted to steadily grow worldwide as it becomes resistant to an increasing number of currently available antibiotics. At the same time, development of new effective antibiotics have dropped precipitously in the past decades. The confluence of these two factors is setting the stage for a "post-antibiotic" era where a great portion of *S. aureus* infections may not be treatable by the existing regimen. In order to stay ahead of this emerging resistance wave, a deeper understanding of the fundamental biology underlying *S. aureus* resistance and pathogenesis is necessary.

Emerging works have demonstrated that resistance and virulence are deeply linked to other aspects of physiology such as metabolism and stress response by the criss-crossing transcriptional regulatory network (TRN). However, untangling these complex regulatory systems from bottom up approaches can be challenging. This dissertation focuses on resolving these complexities in transcriptional regulation by applying Independent Component Analysis (ICA) to RNA sequencing data. ICA recovers the underlying signals from regulators that come together to shape the expression profile of the cell. The result is a scalable, interpretable and functional model of the TRN. We utilized ICA to describe the structure and composition of the TRN in *S. aureus* USA300 strains. Next, we used the TRN model in conjunction with metabolic models to understand how metabolic and regulatory cross-talks coordinate carbon and nitrogen metabolism and direct protein synthesis. Finally, we modeled TRNs from multiple strains and revealed how gene-regulator interactions have evolved during the emergence of the endemic USA300 lineage. Together, this demonstrates the utility of ICA in studying the TRN of *S. aureus* to rapidly discover its structure, dyanmics and evolution over time.

Chapter 1

Introduction

The transcriptional regulatory networks (TRN) control gene expression in response to cues and signals from the environment. In bacteria, this network can consist of transcription factors, sigma factors, riboswitches, regulatory RNAs and various other regulatory elements that work together to shape the final expression profile of the cell[1]. Though simple in comparison to eukaryotes, bacterial TRN with fewer elements can still be difficult to unravel. Even in *Escherichia coli*, perhaps the most well studied bacteria, a significant subset of regulators are still yet to be characterized and our understanding of the TRN in other non-model organisms is sparser still[2, 3]. As the TRN acts as an interface between environmental changes and the subsequent physiological response, understanding the TRN can reveal the mechanisms by which bacteria are able to adapt to challenges and constraints presented by the changing environments. This work focuses on modeling the TRN of *Staphylococcus aureus* USA300 strain using Independent Component Analysis (ICA). The ensuing model was used to understand the structure, dynamics and evolution of the TRN in this important human pathogen.

1.1 Community Associated Methicillin Resistant *S. aureus*

S. aureus causes a variety of human diseases ranging from skin and soft tissue infections (SSTI) to infective endocarditis and pneumonia[4]. This pathogen can also thrive as part of the commensal microbiome in the anterior nares of healthy patients[5]. Rapidly growing resistance to large swaths of antibiotics in *S. aureus* is especially troublesome, earning it a 'high priority' pathogen designation by the World Health Organization[6]. In 2019 alone, more than 700,000 deaths worldwide were attributable to antimicrobial resistance (AMR) associated with *S. aureus*[7], while the deaths attributable to AMR in all pathogens is projected to be the largest cause of mortality by 2050[8].

Different strains of *S. aureus* are endemic to different regions of the world, and in the United States, Community-Associated Methicillin Resistant *S.aureus* (CA-MRSA) USA300 lineage from Clonal Complex 8 (CC8) have become the dominant resistant strain over the past two decades[9]. As their description suggests, these strains harboring AMR genes, are no longer confined to nosocomial environments and can spread rapidly within the community. At the genetic level, this clinical success of USA300 strain has been largely attributed to several horizontally acquired genetic elements including Staphylococcal Chromosomal Cassette mec (SCCMec) harboring beta-lactam resistance gene *mecA*, Panton Valentine Leukocidin (PVL) carrying prophage, Arginine Catabolite Mobile Element (ACME) etc[10]. Beyond the phenotypes that can be directly attributed to mobile genetic elements, differences in toxin expression[11, 12], metabolism[13], virulence regulation[14], biofilm formation[15], and colonization sites [16, 17] have also been observed in USA300. However, the underlying mechanisms that lead to these clinically relevant phenotypes and their interactions with other aspects of *S.aureus* biology remains difficult to untangle. In this dissertation, we focus on gene regulation and its role in the manifestation of these and other clinically important phenotypes.

1.1.1 The transcriptional regulatory network of *S. aureus*

At just over 2.8 million base-pairs, the *S. aureus* USA300 genome contains roughly 2900 total genes and is estimated to contain 135 transcription factors and sigma factors[18]. This consists of 16 two-component systems (TCS), 4 sigma factors and 115 transcription factors. In addition to these proteins, *S. aureus* also encodes various riboswitches, regulatory RNAs, attenuators, t-boxes and other regulatory elements that control transcription. Working in concert with transcription and sigma factors, these regulatory elements are known to control metabolic genes [19, 20], virulence factors [21, 22], biofilm formation [23], and stress response [24]. Together, these regulators take in signals and cues from the environment and alter the physiological state of the cell by changing gene expression levels.

For almost half of the predicted transcription factors, neither the function nor the signals or cues they respond to are currently known[18]. Even for many of the well studied transcription factors, there are still key pieces of information missing. For example, signal activating the closely studied ArlRS two component system which controls virulence expression and biofilm formation, is still unknown [25]. And new roles for transcription factors and signaling pathways that control flux through deeply conserved central metabolism are still coming into view[26, 27]. Additionally, the large number of possible interactions among these regulators in the forms of cross-reactivity[28], signaling cascades[25, 29] or co-regulation[20, 30] are yet to be explored and are the focus of current research in *S.aureus.* Despite these current limitations in our understanding of the gene regulation in *S.aureus*, rapid developments are bringing their importance in clinical settings to light.

Observed changes in gene expression patterns and gene regulation in clinical strains suggests altering these response mechanisms plays an important role in *S.aureus* evolution. The evolution of gene regulation is currently best understood in the context of virulence regulation by Agr, a quorum sensing two component system[31]. In addition to horizontally acquiring PVL toxins, USA300 and other clinical strains also show higher expression levels genes *hla*, *hld*, and *psm* which encode various toxins[12, 32]. As these genes are regulated by the Agr system, high expression of these toxins is attributed to an 'overactive' Agr activity in these strains and have been proposed to be one of the keys to the strains' clinical success[10, 33]. Paradoxically, mutationally inactive Agr systems are frequently isolated from clinical samples[34–36], and have been associated with higher mortality rates in patients [37]. Recently, *Gor et al.* demonstrated that some of the Agr inactivating mutations can revert and restore Agr activity, suggesting that inactivation is a phase variation that enables *S.aureus* to quickly adapt to different environments[38]. While the mechanisms involved and its significance is yet to be sorted, species-wide comparative genome analysis has found that frameshift mutations are common in Agr genes and may point to a convergent evolution in the clinic[39].

Beyond Agr, mutations in other regulators are often observed in response to stress. Mutations altering the activity of pyrimidine biosynthesis regulator, PyrR, and a gene encoding repressor of surface protein, RSP, have been enriched in strains isolated from patients[13, 21]. In laboratory settings, mutations in purine biosynthesis regulator, PurR, can emerge as response to prolonged stress and mutations in VraRST, a cell wall associated regulatory system, are found in strains with increased vancomycin resistance[40–42]. These observations underlie the importance that regulatory adaptations play in *S.aureus* pathogenesis and antibiotic resistance. Chapter 4 explores how these regulatory adaptations can be modeled and studied in emerging clinical strains.

1.1.2 Transcriptional regulation of central metabolic pathways by CcpA and CodY

While the regulation of metabolism in *S.aureus* depends on a large array of regulators working on multiple levels, here we focus on two global regulators, CcpA and CodY, that play critical roles in regulating genes in central carbon and nitrogen metabolism respectively.

CcpA is one of the major carbon catabolite repressors in low-GC gram-positive bacteria[43]. In the presence of favorable carbon sources in the environment such as glucose, CcpA represses the expression of genes involved in alternate carbon uptake and catabolism. Repression by CcpA is initiated when high fructose-1,6-bisphosphate concentration in the cell leads HPr kinase to phosphorylate HPr at Ser-46. When phosphorylated at Ser-46 position, HPr forms a complex with CcpA and together they bind to catabolite responsive element (cre) sites to block gene expression[44]. While CcpA acts as a repressor for most genes in its regulon, some genes can be activated by it in a process known as catabolite activation[45].

CcpA is primarily thought to regulate central metabolic pathways and carbon catabolism, though it also plays a role in controlling other aspects of *S.aureus* physiology. In strain COL, knocking out CcpA led to decrease in alpha toxin production and reduced the minimum inhibitory concentration (MIC) of oxacillin[46]. *In vivo*, CcpA was also shown to be important in establishing infection in hyperglycemic non-obese diabetic mice model[47]. Coordinating its activity with MgrA, ArlRS, and CidR, CcpA also plays a role in biofilm formation [27, 48]. In line with this expanded role of CcpA in different aspects of *S.aureus* physiology, CcpA activity can be altered when phosphorylated by serine/threonine protein kinase Stk1[27]. Stk1 is a regulator of cell wall biosynthesis and its ability to phosphorylate CcpA provides an entry point for incorporating cell wall stress associated signals into the regulation of central carbon metabolism and virulence[49, 50]. Similarly, CcpA also interacts closely with CodY, a global regulator of amino acid metabolism in firmicutes, to coordinate central carbon metabolism with nitrogen metabolism[51].

Like CcpA, Cody is also a global metabolic regulator, but primarily regulates the biosynthesis of amino acids instead[52]. CodY responds to cellular concentrations of branched chain amino acids (isoleucine, leucine and valine) and GTP[53]. Though concentrations of any of these effectors can additively change CodY activity, isoleucine concentration seems to be the dominant cue in *S.aureus*[20]. In the presence of high concentrations of its effectors, CodY represses the expression of various genes involved in amino acid biosynthesis, capsule proteins, lipoproteins, and peptide and ion transporters[54].

CodY, similar to CcpA, also has an expanded role in connecting regulation of metabolism to other aspects of *S.aureus* physiology. The role of CodY as a "regulatory link between metabolism and virulence" has now been well characterized[55, 56]. It plays an important role in modulating toxin expression via Agr, can regulate polysaccharide intercellular adhesin (PIA), and represses the expression of *nuc* which encodes neutrophil extracellular traps degrading nuclease[52]. In addition to regulating virulence factors, CodY is also an integral part of stringent response in gram-positive bacteria[57]. Stringent response activates in the presence of environmental stress such as nutrient deprivation, when RelA/SpoT homolog (RSH) rapidly converts GTP to signaling molecule (p)ppgpp [58]. Though (p)pgpp can independently lead to changes in gene expression and physiology [59], synthesis of (p)pgpp also leads to drop in cellular GTP level and therefore causes CodY derepression[57, 60]. The synthesis of (p)pgpp by RelA is coupled to ribosome stalling[61], but the alarmone can also be synthesized by two other homologs, RelP and RelQ[62]. Genes encoding RelP and RelQ are part of VraR regulon and deletion of these enzymes leads to lower tolerance for cell wall stress[63]. Similar to RelA, RelQ can also lead to CodY activation[64], suggesting that RelP and RelQ can affect CodY activity in response to cell wall stress.

CcpA and Cody activity regulate carbon and nitrogen metabolism respectively in *S.aureus*. Their roles however expand beyond just metabolic regulation as they can both affect virulence expression, antibiotic resistance levels and infectivity. With this expanded role, both regulators also seem to integrate signals from stress response pathways into its activities thereby modulating the central carbon and nitrogen metabolism in response to environmental stresses. In chapter 3, we expand on the activities of these two regulators and how they are coordinated to regulate various stages of protein synthesis.

1.2 Modeling gene regulation with Independent Component Analysis

Independent Component Analysis (ICA) is blind source separation algorithm designed to extract source signals from a mixture of unknown signals[65]. The source signals can be sounds, signals from radio, readings of brain activity from electroencephalograph, or in our case signals from gene regulators. The key insight in finding these signals from a mixture is provided by the Central Limit Theorem (CLT), which states that the sum of independent and identically distributed non-gaussian variables tends to be closer to gaussian than the input[66]. In other words, when non-gaussian signals are mixed together, the resulting mixed signal tends to be more gaussian than the input signals. Therefore, to find the source signals ICA searches for components along which non-gaussianity is maximized. These components can be discovered using an efficient fixed-point algorithm named FastICA[67].

We can apply FastICA to a compendium of RNA sequencing data to extract the source regulatory signals[68]. The compendium, referred to as the expression ($\boldsymbol{X}_{genes,samples}$) matrix, contains all publicly available RNA sequencing data for USA300 strains that have been pulled from Sequence Read Archive (SRA). ICA then factorizes the $\boldsymbol{X}$ matrix into a modulon ($\boldsymbol{M}_{genes,components}$) matrix and an activity ($\boldsymbol{A}_{components,samples}$) matrix. Each column of the **M**-matrix contains weighting for all the genes in the input RNA sequencing data and represents one of the linearly independent source signals from which independently modulated sets of genes (iModulons) can be retrieved. In other words, within each column of the $\boldsymbol{M}$-matrix, there can be found a set of genes whose expression level is resultant from a particular independent singal. The $\boldsymbol{A}$-matrix then contains the activity of each of these signals in all the input RNA sequencing samples. Together, the two matrices contain information about the set of genes that are regulated together (iModulons) and how the regulators act in each of the samples, thus providing both the sample-invariant and the sample dependent aspects of the TRN.

This framework for determining the TRN is ideal for non-model organisms as it generates a scalable, interpretable and functional model. As the TRN is calculated from the exponentially growing collections of public RNA sequencing data, the model grows with each new profile generated by the community. Due to its scalability, ICA has been used to quickly generate models for seven bacterial species and even an archea and will continue to grow in size over time[69]. The model also provides interpretable view of the TRN, giving insights into both its static and dynamic aspects. Finally, each model is functional, enabling a wide array of various analyses including interpretation complex expression profiles, integration with various other types of models and comparative analysis of the evolution of gene networks across the phylogenetic tree.

1.3 Dissertation Outline

In this work, we apply ICA to RNA sequencing data from *S. aureus* USA300 strain to generate a model of its TRN. Next chapter describes the first model for this strain and its uses in describing new gene-regulator relationships, interpreting complex *in vivo* expression profiles and defining new global interactions among regulators. In the third chapter, we integrate ICA and metabolic models to understand the dynamics of two regulators, CcpA and CodY, and how they coordinate their activities to control protein production. Finally, the fourth chapter uses the model to extend genome wide association studies (GWAS) and predict the evolution of the regulatory network during the emergence of the endemic USA300 strains. Together, ICA is used to define the structure, dynamics and evolution of TRN in the USA300 strains.

1.4 References

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

2. Gao, Y., Yurkovich, J. T., Seo, S. W., Kabimoldayev, I., Dräger, A., Chen, K., Sastry, A. V., Fang, X., Mih, N., Yang, L., Eichner, J., Cho, B.-K., Kim, D. & Palsson, B. O. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. en. *Nucleic Acids Res.* **46,** 10682–10696 (Nov. 2018).

3. Rodionova, I. A., Gao, Y., Sastry, A., Hefner, Y., Lim, H. G., Rodionov, D. A., Saier Jr, M. H. & Palsson, B. O. Identification of a transcription factor, PunR, that regulates the purine and purine nucleoside transporter punC in E. coli. en. *Commun Biol* **4,** 991 (Aug. 2021).

4. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L. & Fowler, V. G. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28,** 603–661 (July 2015).

5. Krismer, B., Weidenmaier, C., Zipperer, A. & Peschel, A. The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. en. *Nat. Rev. Microbiol.* **15,** 675–687 (Oct. 2017).

6. Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E. M., Houchens, C. R., Grayson, M. L., Hansen, P., Singh, N., Theuretzbacher, U., Magrini, N. & WHO Pathogens Priority List Working Group. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. en. *Lancet Infect. Dis.* **18,** 318–327 (Mar. 2018).

7. Murray, C. J. L. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. en. *Lancet* **399,** 629–655 (Feb. 2022).

8. Review on Antimicrobial Resistance (London). & Grande-Bretagne. *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations* en (Review on Antimicrobial Resistance, 2014).

9. Planet, P. J. Life After USA300: The Rise and Fall of a Superbug. en. *J. Infect. Dis.* **215,** S71–S77 (Feb. 2017).

10. Thurlow, L. R., Joshi, G. S. & Richardson, A. R. Virulence strategies of the dominant USA300 lineage of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA). en. *FEMS Immunol. Med. Microbiol.* **65,** 5–22 (June 2012).

11. Kobayashi, S. D., Malachowa, N., Whitney, A. R., Braughton, K. R., Gardner, D. J., Long, D., Bubeck Wardenburg, J., Schneewind, O., Otto, M. & Deleo, F. R. Comparative analysis of USA300 virulence determinants in a rabbit model of skin and soft tissue infection. en. *J. Infect. Dis.* **204,** 937–941 (Sept. 2011).

12. Kennedy, A. D., Otto, M., Braughton, K. R., Whitney, A. R., Chen, L., Mathema, B., Mediavilla, J. R., Byrne, K. A., Parkins, L. D., Tenover, F. C., Kreiswirth, B. N., Musser, J. M. & DeLeo, F. R. Epidemic community-associated methicillin-resistant Staphylococcus aureus: recent clonal expansion and diversification. en. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1327–1332 (Jan. 2008).

13. Copin, R., Sause, W. E., Fulmer, Y., Balasubramanian, D., Dyzenhaus, S., Ahmed, J. M., Kumar, K., Lees, J., Stachel, A., Fisher, J. C., Drlica, K., Phillips, M., Weiser, J. N., Planet, P. J., Uhlemann, A.-C., Altman, D. R., Sebra, R., van Bakel, H., Lighter, J., Torres, V. J. & Shopsin, B. Sequential evolution of virulence and resistance during clonal spread of community-acquired methicillin-resistant Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 1745–1754 (Jan. 2019).

14. Sause, W. E., Balasubramanian, D., Irnov, I., Copin, R., Sullivan, M. J., Sommerfield, A., Chan, R., Dhabaria, A., Askenazi, M., Ueberheide, B., Shopsin, B., van Bakel, H. &

Torres, V. J. The purine biosynthesis regulator PurR moonlights as a virulence regulator in Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 13563–13572 (July 2019).

15. Vanhommerig, E., Moons, P., Pirici, D., Lammens, C., Hernalsteens, J.-P., De Greve, H., Kumar-Singh, S., Goossens, H. & Malhotra-Kumar, S. Comparison of biofilm formation between major clonal lineages of methicillin resistant Staphylococcus aureus. en. *PLoS One* **9,** e104561 (Aug. 2014).

16. Faden, H., Lesse, A. J., Trask, J., Hill, J. A., Hess, D. J., Dryja, D. & Lee, Y.-H. Importance of colonization site in the current epidemic of staphylococcal skin abscesses. en. *Pediatrics* **125,** e618–24 (Mar. 2010).

17. McCullough, A. C., Seifried, M., Zhao, X., Haase, J., Kabat, W. J., Yogev, R., Blumenthal, R. M. & Mukundan, D. Higher incidence of perineal community acquired MRSA infections among toddlers. en. *BMC Pediatr.* **11,** 96 (Oct. 2011).

18. Ibarra, J. A., Pérez-Rueda, E., Carroll, R. K. & Shaw, L. N. Global analysis of transcriptional regulators in Staphylococcus aureus. en. *BMC Genomics* **14,** 126 (Feb. 2013).

19. Lünse, C. E., Schmidt, M. S., Wittmann, V. & Mayer, G. Carba-sugars activate the glmS-riboswitch of Staphylococcus aureus. en. *ACS Chem. Biol.* **6,** 675–678 (July 2011).

20. Kaiser, J. C., King, A. N., Grigg, J. C., Sheldon, J. R., Edgell, D. R., Murphy, M. E. P., Brinsmade, S. R. & Heinrichs, D. E. Repression of branched-chain amino acid synthesis in Staphylococcus aureus is mediated by isoleucine via CodY, and by a leucine-rich attenuator peptide. en. *PLoS Genet.* **14,** e1007159 (Jan. 2018).

21. Das, S., Lindemann, C., Young, B. C., Muller, J., Österreich, B., Ternette, N., Winkler, A.-C., Paprotka, K., Reinhardt, R., Förstner, K. U., Allen, E., Flaxman, A., Yamaguchi, Y., Rollier, C. S., van Diemen, P., Blättner, S., Remmele, C. W., Selle, M., Dittrich, M., Müller, T., Vogel, J., Ohlsen, K., Crook, D. W., Massey, R., Wilson, D. J., Rudel, T., Wyllie, D. H. & Fraunholz, M. J. Natural mutations in a Staphylococcus aureus virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3101–10 (May 2016).

22. Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A. C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F. & Romby, P. Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. en. *Genes Dev.* **21,** 1353–1366 (June 2007).

23. Romilly, C., Lays, C., Tomasini, A., Caldelari, I., Benito, Y., Hammann, P., Geissmann, T., Boisset, S., Romby, P. & Vandenesch, F. A non-coding RNA promotes bacterial persistence and decreases virulence by regulating a regulator in Staphylococcus aureus. en. *PLoS Pathog.* **10,** e1003979 (Mar. 2014).

24. Augagneur, Y., King, A. N., Germain-Amiot, N., Sassi, M., Fitzgerald, J. W., Sahukhal, G. S., Elasri, M. O., Felden, B. & Brinsmade, S. R. Analysis of the CodY RNome reveals

RsaD as a stress-responsive riboregulator of overflow metabolism in Staphylococcus aureus. en. *Mol. Microbiol.* **113,** 309–325 (Feb. 2020).

25. Crosby, H. A., Tiwari, N., Kwiecinski, J. M., Xu, Z., Dykstra, A., Jenul, C., Fuentes, E. J. & Horswill, A. R. The Staphylococcus aureus ArlRS two-component system regulates virulence factor expression through MgrA. en. *Mol. Microbiol.* **113,** 103–122 (Jan. 2020).

26. Ding, Y., Liu, X., Chen, F., Di, H., Xu, B., Zhou, L., Deng, X., Wu, M., Yang, C.-G. & Lan, L. Metabolic sensor governing bacterial virulence in Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E4981–90 (Nov. 2014).

27. Leiba, J., Hartmann, T., Cluzel, M.-E., Cohen-Gonsaud, M., Delolme, F., Bischoff, M. & Molle, V. A Novel Mode of Regulation of the *Staphylococcus aureus* Catabolite Control Protein A (CcpA) Mediated by Stk1 Protein Phosphorylation. *J. Biol. Chem.* **287,** 43607–43619 (Dec. 2012).

28. Villanueva, M., Garcıa, B., Valle, J., Rapún, B., Ruiz de Los Mozos, I., Solano, C., Martı, M., Penadés, J. R., Toledo-Arana, A. & Lasa, I. Sensory deprivation in Staphylococcus aureus. en. *Nat. Commun.* **9,** 523 (Feb. 2018).

29. Kwiecinski, J. M., Kratofil, R. M., Parlet, C. P., Surewaard, B. G. J., Kubes, P. & Horswill, A. R. Staphylococcus aureus uses the ArlRS and MgrA cascade to regulate immune evasion during skin infection. en. *Cell Rep.* **36,** 109462 (July 2021).

30. Reed, J. M., Olson, S., Brees, D. F., Griffin, C. E., Grove, R. A., Davis, P. J., Kachman, S. D., Adamec, J. & Somerville, G. A. Coordinated regulation of transcription by CcpA and the Staphylococcus aureus two-component system HptRS. en. *PLoS One* **13,** e0207161 (Dec. 2018).

31. Yarwood, J. M. & Schlievert, P. M. Quorum sensing in Staphylococcus infections. en. *J. Clin. Invest.* **112,** 1620–1625 (Dec. 2003).

32. Wang, R., Braughton, K. R., Kretschmer, D., Bach, T.-H. L., Queck, S. Y., Li, M., Kennedy, A. D., Dorward, D. W., Klebanoff, S. J., Peschel, A., DeLeo, F. R. & Otto, M. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. en. *Nat. Med.* **13,** 1510–1514 (Dec. 2007).

33. Li, M., Cheung, G. Y. C., Hu, J., Wang, D., Joo, H.-S., Deleo, F. R. & Otto, M. Comparative analysis of virulence and toxin expression of global community-associated methicillin-resistant Staphylococcus aureus strains. en. *J. Infect. Dis.* **202,** 1866–1876 (Dec. 2010).

34. Suligoy, C. M., Lattar, S. M., Noto Llana, M., González, C. D., Alvarez, L. P., Robinson, D. A., Gómez, M. I., Buzzola, F. R. & Sordelli, D. O. Mutation of Agr Is Associated with the Adaptation of Staphylococcus aureus to the Host during Chronic Osteomyelitis. en. *Front. Cell. Infect. Microbiol.* **8,** 18 (Feb. 2018).

35. Shopsin, B., Drlica-Wagner, A., Mathema, B., Adhikari, R. P., Kreiswirth, B. N. & Novick, R. P. Prevalence of agr dysfunction among colonizing Staphylococcus aureus strains. en. *J. Infect. Dis.* **198,** 1171–1174 (Oct. 2008).

36. Traber, K. E., Lee, E., Benson, S., Corrigan, R., Cantera, M., Shopsin, B. & Novick, R. P. agr function in clinical Staphylococcus aureus isolates. en. *Microbiology* **154,** 2265–2274 (Aug. 2008).

37. Schweizer, M. L., Furuno, J. P., Sakoulas, G., Johnson, J. K., Harris, A. D., Shardell, M. D., McGregor, J. C., Thom, K. A. & Perencevich, E. N. Increased mortality with accessory gene regulator (agr) dysfunction in Staphylococcus aureus among bacteremic patients. en. *Antimicrob. Agents Chemother.* **55,** 1082–1087 (Mar. 2011).

38. Gor, V., Takemura, A. J., Nishitani, M., Higashide, M., Medrano Romero, V., Ohniwa, R. L. & Morikawa, K. Finding of Agr Phase Variants in Staphylococcus aureus. en. *MBio* **10** (Aug. 2019).

39. Raghuram, V., Alexander, A. M., Loo, H. Q., Petit 3rd, R. A., Goldberg, J. B. & Read, T. D. Species-Wide Phylogenomics of the Staphylococcus aureus Agr Operon Revealed Convergent Evolution of Frameshift Mutations. en. *Microbiol Spectr,* e0133421 (Jan. 2022).

40. Goncheva, M. I., Flannagan, R. S., Sterling, B. E., Laakso, H. A., Friedrich, N. C., Kaiser, J. C., Watson, D. W., Wilson, C. H., Sheldon, J. R., McGavin, M. J., Kiser, P. K. & Heinrichs, D. E. Stress-induced inactivation of the Staphylococcus aureus purine biosynthesis repressor leads to hypervirulence. en. *Nat. Commun.* **10,** 775 (Feb. 2019).

41. Machado, H., Seif, Y., Sakoulas, G., Olson, C. A., Hefner, Y., Anand, A., Jones, Y. Z., Szubin, R., Palsson, B. O., Nizet, V. & Feist, A. M. Environmental conditions dictate differential evolution of vancomycin resistance in Staphylococcus aureus. en. *Commun Biol* **4,** 793 (June 2021).

42. Kato, Y., Suzuki, T., Ida, T. & Maebashi, K. Genetic changes associated with glycopeptide resistance in Staphylococcus aureus: predominance of amino acid substitutions in YvqF/VraSR. en. *J. Antimicrob. Chemother.* **65,** 37–45 (Jan. 2010).

43. Titgemeyer, F. & Hillen, W. Global control of sugar metabolism: a gram-positive solution. en. *Antonie Van Leeuwenhoek* **82,** 59–71 (Aug. 2002).

44. Lorca, G. L., Chung, Y. J., Barabote, R. D., Weyler, W., Schilling, C. H. & Saier Jr, M. H. Catabolite repression and activation in Bacillus subtilis: dependency on CcpA, HPr, and HprK. en. *J. Bacteriol.* **187,** 7826–7839 (Nov. 2005).

45. Seidl, K., Müller, S., François, P., Kriebitzsch, C., Schrenzel, J., Engelmann, S., Bischoff, M. & Berger-Bächi, B. Effect of a glucose impulse on the CcpA regulon in Staphylococcus aureus. en. *BMC Microbiol.* **9,** 95 (May 2009).

46. Seidl, K., Stucki, M., Ruegg, M., Goerke, C., Wolz, C., Harris, L., Berger-Bächi, B. & Bischoff, M. Staphylococcus aureus CcpA affects virulence determinant production and antibiotic resistance. en. *Antimicrob. Agents Chemother.* **50,** 1183–1194 (Apr. 2006).

47. Bischoff, M., Wonnenberg, B., Nippe, N., Nyffenegger-Jann, N. J., Voss, M., Beisswenger, C., Sunderkötter, C., Molle, V., Dinh, Q. T., Lammert, F., Bals, R., Herrmann, M., Somerville,

G. A., Tschernig, T. & Gaupp, R. CcpA Affects Infectivity of Staphylococcus aureus in a Hyperglycemic Environment. en. *Front. Cell. Infect. Microbiol.* **7,** 172 (May 2017).

48. Sadykov, M. R., Windham, I. H., Widhelm, T. J., Yajjala, V. K., Watson, S. M., Endres, J. L., Bavari, A. I., Thomas, V. C., Bose, J. L. & Bayles, K. W. CidR and CcpA Synergistically Regulate Staphylococcus aureus cidABC Expression. en. *J. Bacteriol.* **201** (Dec. 2019).

49. Beltramini, A. M., Mukhopadhyay, C. D. & Pancholi, V. Modulation of cell wall structure and antimicrobial susceptibility by a Staphylococcus aureus eukaryote-like serine/threonine kinase and phosphatase. en. *Infect. Immun.* **77,** 1406–1416 (Apr. 2009).

50. Jarick, M., Bertsche, U., Stahl, M., Schultz, D., Methling, K., Lalk, M., Stigloher, C., Steger, M., Schlosser, A. & Ohlsen, K. The serine/threonine kinase Stk and the phosphatase Stp regulate cell wall synthesis in Staphylococcus aureus. en. *Sci. Rep.* **8,** 13693 (Sept. 2018).

51. Sonenshein, A. L. Control of key metabolic intersections in Bacillus subtilis. en. *Nat. Rev. Microbiol.* **5,** 917–927 (Dec. 2007).

52. Majerczyk, C. D., Dunman, P. M., Luong, T. T., Lee, C. Y., Sadykov, M. R., Somerville, G. A., Bodi, K. & Sonenshein, A. L. Direct targets of CodY in Staphylococcus aureus. en. *J. Bacteriol.* **192,** 2861–2877 (June 2010).

53. Handke, L. D., Shivers, R. P. & Sonenshein, A. L. Interaction of Bacillus subtilis CodY with GTP. en. *J. Bacteriol.* **190,** 798–806 (Feb. 2008).

54. Waters, N. R., Samuels, D. J., Behera, R. K., Livny, J., Rhee, K. Y., Sadykov, M. R. & Brinsmade, S. R. A spectrum of CodY activities drives metabolic reorganization and virulence gene expression in Staphylococcus aureus. en. *Mol. Microbiol.* **101,** 495–514 (Aug. 2016).

55. Pohl, K., Francois, P., Stenz, L., Schlink, F., Geiger, T., Herbert, S., Goerke, C., Schrenzel, J. & Wolz, C. CodY in Staphylococcus aureus: a regulatory link between metabolism and virulence gene expression. en. *J. Bacteriol.* **191,** 2953–2963 (May 2009).

56. Majerczyk, C. D., Sadykov, M. R., Luong, T. T., Lee, C., Somerville, G. A. & Sonenshein, A. L. Staphylococcus aureus CodY negatively regulates virulence gene expression. en. *J. Bacteriol.* **190,** 2257–2265 (Apr. 2008).

57. Intersection of the stringent response and the CodY regulon in low GC Gram-positive bacteria. *Int. J. Med. Microbiol.* **304,** 150–155 (Mar. 2014).

58. Wendrich, T. M. & Marahiel, M. A. Cloning and characterization of a relA/spoT homologue from Bacillus subtilis. en. *Mol. Microbiol.* **26,** 65–79 (Oct. 1997).

59. Potrykus, K. & Cashel, M. (p)ppGpp: still magical? en. *Annu. Rev. Microbiol.* **62,** 35–51 (2008).

60. Geiger, T., Francois, P., Liebeke, M., Fraunholz, M., Goerke, C., Krismer, B., Schrenzel, J., Lalk, M. & Wolz, C. The stringent response of Staphylococcus aureus and its impact

on survival after phagocytosis through the induction of intracellular PSMs expression. en. *PLoS Pathog.* **8,** e1003016 (Nov. 2012).

61. Loveland, A. B., Bah, E., Madireddy, R., Zhang, Y., Brilot, A. F., Grigorieff, N. & Korostelev, A. A. Ribosome•RelA structures reveal the mechanism of stringent response activation. en. *Elife* **5** (July 2016).

62. Lemos, J. A., Lin, V. K., Nascimento, M. M., Abranches, J. & Burne, R. A. Three gene products govern (p)ppGpp production by Streptococcus mutans. en. *Mol. Microbiol.* **65,** 1568–1581 (Sept. 2007).

63. Geiger, T., Kästle, B., Gratani, F. L., Goerke, C. & Wolz, C. Two small (p)ppGpp synthases in Staphylococcus aureus mediate tolerance against cell envelope stress conditions. en. *J. Bacteriol.* **196,** 894–902 (Feb. 2014).

64. Horvatek, P., Salzer, A., Hanna, A. M. F., Gratani, F. L., Keinhörster, D., Korn, N., Borisova, M., Mayer, C., Rejman, D., Mäder, U. & Wolz, C. Inducible expression of (pp)pGpp synthetases in Staphylococcus aureus is associated with activation of stress response genes. en. *PLoS Genet.* **16,** e1009282 (Dec. 2020).

65. Hyvarinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* (JOHN WILEY & SONS, INC, Mar. 2001).

66. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36,** 287–314 (Apr. 1994).

67. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. en. *Neural Netw.* **13,** 411–430 (May 2000).

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

69. Rychel, K., Decker, K., Sastry, A. V., Phaneuf, P. V., Poudel, S. & Palsson, B. O. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. en. *Nucleic Acids Res.* **49,** D112–D120 (Jan. 2021).

Chapter 2

Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response

2.1 Abstract

The ability of *Staphylococcus aureus* to infect many different tissue sites is enabled, in part, by its Transcriptional Regulatory Network (TRN) that coordinates its gene expression to respond to different environments. We elucidated the organization and activity of this TRN by applying Independent Component Analysis (ICA) to a compendium of 108 RNAseq expression profiles from two *S. aureus* clinical strains (TCH1516 and LAC). ICA decomposed the *S. aureus*

transcriptome into 29 independently modulated sets of genes (iModulons) that revealed (1) high confidence associations between 21 iModulons and known regulators; (2) an association between an iModulon and σS, whose regulatory role was previously undefined; (3) the regulatory organization of 65 virulence factors in the form of three iModulons associated with AgrR, SaeR and Vim-3, (4) the roles of three key transcription factors (CodY, Fur and CcpA) in coordinating the metabolic and regulatory networks; and (5) a low dimensional representation, involving the function of few transcription factors, of changes in gene expression between two laboratory media (RPMI, CAMHB) and two physiological media (blood and serum). This representation of the TRN covers 842 genes representing 76% of the variance in gene expression that provides a quantitative reconstruction of transcriptional modules in *S. aureus*, and a platform enabling its full elucidation.

2.2 Introduction

Staphylococcus aureus causes a variety of human diseases ranging from skin and soft tissue infections (SSTI) to infective endocarditis and pneumonia[1]. The pathogen can also thrive as part of the commensal microbiome in the anterior nares of healthy patients[2]. *S. aureus* adaptation to many different host environments is enabled, in part, by the underlying transcriptional regulatory network (TRN) that can alter the physiological state of the cell to match the unique challenges presented by each environment[3–5]. Such adaptations require coordinated expression of genes in many cellular subsystems such as metabolism, cell wall biosynthesis, stress response, virulence factors, etc. Therefore, a complete understanding of the *S. aureus* response to different environments necessitates a thorough understanding of its TRN. However, since *S. aureus* is predicted to have as many as 135 transcriptional regulators[6], with many more potential interactions among them, a bottom-up study of its global TRN becomes intractable.

To address this challenge, we previously introduced an Independent Component Analysis (ICA)-based framework in *Escherichia coli* that decomposes a compendium of RNA-sequencing (RNA-seq) expression profiles to determine the underlying regulatory structure[7]. An extensive analysis of module detection methods demonstrated that ICA out-performed most other methods in consistently recovering known biological modules[8]. The framework defines independently modulated sets of genes (called iModulons) and calculates the activity level of each iModulon in the input expression profile. ICA analysis of expression profiles in *E. coli* have been used to describe undefined regulons, link strain-specific mutations with changes in gene expression, and understand rewiring of TRN during Adaptive Laboratory Evolution (ALE)[7, 9]. Given the deeper insights it provided into the TRN of *E. coli*, we sought to expand this approach to the human pathogen *S. aureus*. To elucidate the TRN features in *S. aureus*, we compiled 108 high quality RNA-seq expression profiles for community-associated methicillin-resistant *S. aureus* (CA-MRSA) strains LAC and TCH1516. Decomposition of these expression profiles revealed 29 independently modulated sets of genes and their activity levels across all 108 expression profiles. Further, we show that using the new framework to reevaluate the RNA-seq data accelerates discovery by (1) quantitatively formulating TRN organization, (2) simplifying complex changes across hundreds of genes into a few changes in regulator activities, (3) allowing for analysis of interactions among different regulators, (4) connecting transcriptional regulation to metabolism, and (5) defining previously unknown regulons.

2.3 Results

2.3.1 ICA extracts biologically meaningful components from transcriptomic data

We generated 108 high-quality RNA-seq expression profiles from CA-MRSA USA300 isolates LAC and TCH1516 and two additional Adaptive Laboratory Evolution (ALE)-derivatives of TCH1516. To capture a wide range of expression states, we collected RNA-seq data from *S. aureus* exposed to various media conditions, antibiotics, nutrient sources, and other stressors. The samples were then filtered for high reproducibility between replicates to minimize noise in the data (Figure A.1a). The final dataset contained 108 samples representing 43 unique growth conditions, which have an average $R2 = 0.98$ between replicates.

Using an extended ICA algorithm[7], we decomposed the expression compendium into 29 'iModulons'. An iModulon contains a set of genes whose expression levels vary concurrently with each other, but independently of all other genes not in the given iModulon. Akin to a regulon[10], an iModulon represents a regulatory organizational unit containing a functionally related and co-expressed set of genes under all conditions considered (Figure 2.1a). While regulons are determined based on direct molecular methods (e.g., ChIP-seq, RIP-ChIP, gene-knockouts, etc.), iModulons are defined through an untargeted ICA-based statistical approach applied to RNA-seq data that is a reflection of the activity of the transcriptional regulators (see Materials and Methods). However, beyond regulons, iModulons can also describe other genomic features, such as strain differences and genetic alterations (e.g. gene knock-out) that can lead to change in gene co-expression[7, 9]. The outcome of this approach is a biologically relevant, low dimensional mathematical representation of functional modules in the TRN that reconstruct most of the

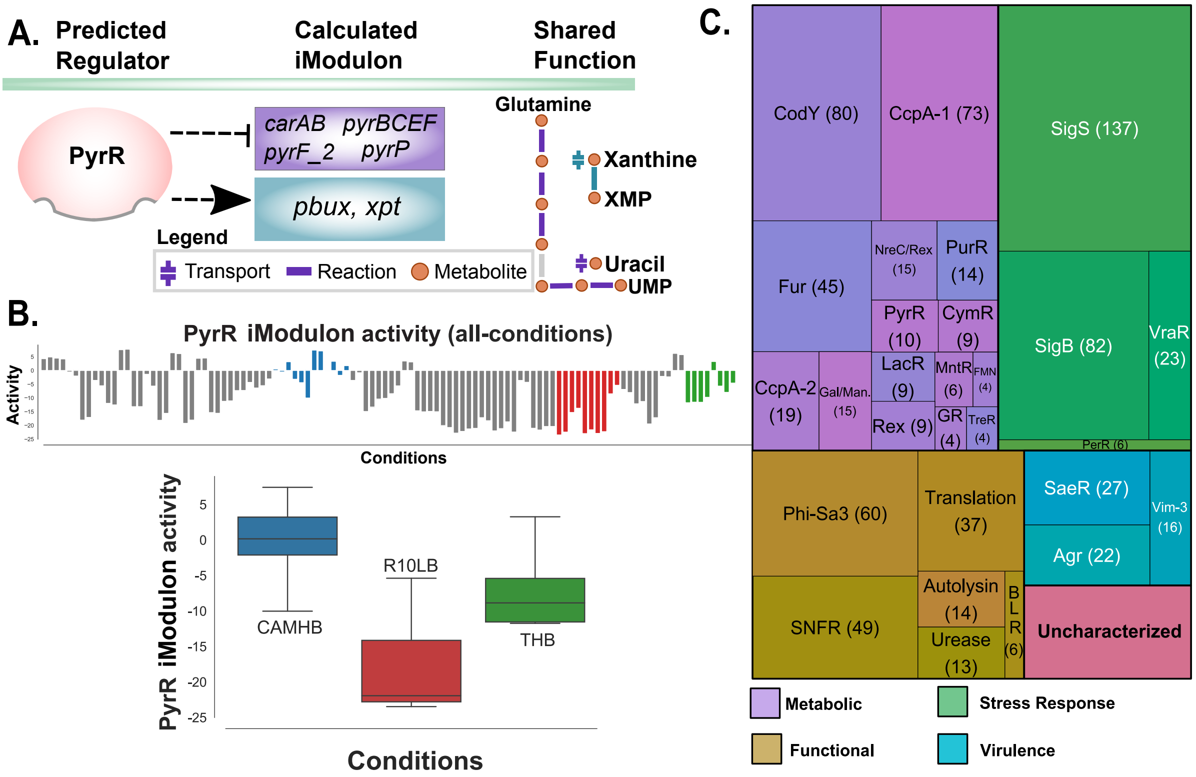


Figure 2.1: ICA decomposition of *S. aureus* USA300 RNA-seq database. (a) An iModulon is a set of genes that are co-expressed and encode products with shared functions. The PyrR iModulon, for example (middle column), is predicted to be under control of pyrR repressor and contains genes that encode enzymes in pyrimidine biosynthesis (purple) and purine salvage (blue) pathway (right column). The genes in two different pathways are contra-regulated (arrows). (b) Activity levels of iModulons are calculated for all conditions (top bar chart), allowing for sample specific (e.g., in three different media) comparison of each iModulon (boxplot). The activity of all iModulons are centered around CAMHB base condition and therefore, all iModulons have mean activity of 0 in this condition. Centerline of the boxplot represents median value, the box limits represent Q1 and Q3, and the whiskers represent the min and max values. (c) A treemap indicating the names and the size of the iModulons. The iModulons are named after the transcription factor(s) whose predicted regulons have highest overlap with the given iModulon, or based on the shared functionality of genes (e.g., autolysins, translation, B-lactam resistance) in iModulon if no known regulator was identified. iModulon with low or no correspondence with any of the known features is labeled as Unc-1. ‘BLR’ stands for ‘Beta-Lactam resistance’ and SNFR iModulon consists of genes with altered expression in SNFR strain.

information content of the input RNAseq compendium (SI Appendix, Figure A.1b). Such formulation also quantitatively captures complex behaviors of regulators such as contra-regulation of multiple genes by the same regulator, co-regulation of the same gene by multiple regulators, and coordinated expression of multiple organizational units (iModulons) in various conditions (SI Appendix, Figure A.1c,d). Therefore, this model enables simultaneous analysis of TRN at both gene and genome-scale.

ICA also reconstructs the activity of the iModulons in the samples, which represents the collective expression level of the genes in the iModulon. Each sample in the dataset can be reconstructed as the summation of the activity of the 29 iModulons, which makes the transcriptional state in each condition more explainable. Conversely, each iModulon has a computed activity in every sample, allowing for easy comparisons of iModulon activities across samples, that in turn reflect the activity of the corresponding transcriptional regulator (Figure 2.1b). The reported activity levels are log2 fold change from the base condition - growth in Cation Adjust Mueller Hinton Broth (CAMHB).

We compared the gene sets in the 29 enriched iModulons against previously predicted *S. aureus* regulons in the RegPrecise database and other regulons described in various publications. iModulons with statistically significant overlap ($FDR < 1e-05$) with a previously predicted regulon were named after the transcription factor associated with the regulon (see Materials and Methods). We also manually identified iModulons that consisted of genes with shared functions (e.g., Autolysins, Translation) or those that corresponded to other genomic features such as plasmids, prophages, or strain-specific differences. Together, we identified fifteen metabolic, six functional, three virulence, four stress response-associated and one strain-associated (SNFR) iModulons (Figure 2.1c). Of the 29 enriched iModulons, only one remains uncharacterized. In total, the 29 iModulons consist of 752 unique genes, 90 of which are enriched in more than 1 iModulons.

2.3.2 ICA disentangles complex change in the transcriptome

Differential expression analysis of *S. aureus* in different environmental conditions can yield hundreds of genes that have significantly altered expression levels, hindering meaningful interpretation. Decomposition of the expression profile into biologically meaningful iModulons instead allows us to gain a comprehensive understanding of the change in the transcriptome through the activities of few regulators. To demonstrate this capability, we explored the difference in expression profiles of *S. aureus* grown in two different media, cation-adjusted Mueller Hinton Broth (CAMHB), the standard bacteriologic medium for routine antimicrobial susceptibility testing worldwide, and the common physiologically relevant mammalian tissue culture medium RPMI-1640, supplemented with 10% Luria broth (RPMI+10%LB) to support growth kinetics similar to CAMHB. Over 800 genes spanning more than a dozen Clusters of Orthologous Groups (COG)[11] categories were differentially expressed between the two media (SI Appendix, Figure A.2a). Conversely, there were fifteen iModulons with statistically significant differential activation (Figure 2.2a). Most differentially activated iModulons were involved in metabolism (CodY, PurR, Guanine-Responsive iModulon (GR), Gal/Man, Rex, MntR, PyrR, LacR, CcpA-1, CcpA-2, Urease). The last four iModulons were those with functions in virulence (Vim-3, SaeR), Translation, and the Phi-Sa3 phage-specific iModulon. Concurrent activation of the CodY, PurR, and GR iModulons in RPMI+10%LB indicates that this media presents a guanine-limited environment, as activity of all three transcription factors decrease in response to falling cellular concentrations of various forms of guanine derivatives[12–15]. Consistent with this hypothesis, we also saw decreased activity of the Translation iModulon in RPMI+10%LB. Downregulation of translation machinery often occurs during the stringent response, where cellular GTP is depleted as it is rapidly converted to ppGpp[12, 16–18]. Similarly, activation of the MntR iModulon points to manganese starvation in RPMI+10%LB, and the decreased activity of two iModulons associated with carbon catabolite repressor CcpA (CcpA-1 and CcpA-2) likely reflects a glucose replete environment[19]. Analysis of spent media using HPLC confirmed that *S. aureus* was actively uptaking glucose in RPMI+10%LB while no glucose was detected in CAMHB (SI Appendix, Figure A.2b). Taken together, the shift in activity of iModulons between the two media suggests that compared to the bacteriologic medium CAMHB, RPMI+10%LB presents an environment poor in purines (specifically guanine) and manganese but rich in the carbon source glucose.

Next, we designed two validation experiments to ensure that the activity level of iModulons reflect expected outcomes. To this end, we chose three iModulons to validate, CcpA-1, CcpA-2, and GR, for ease of modifying their activities with supplementation of glucose and purines, respectively. CcpA is the carbon catabolite repressor in *S. aureus* that controls central carbon metabolism and carbon source utilization[20, 21]. Its activity level is indirectly modulated by cellular glucose concentration, though it can also be altered by other glucose-independent signals[22, 23]. CcpA transcriptional effects are captured in two iModulons, CcpA-1 and CcpA-2, which contain 73 and 19 genes, respectively. Both iModulons had far lower activity in RPMI+10%LB compared to CAMHB. However, the addition of 2g/L glucose only led to reduced activity of the CcpA-1 iModulon in CAMHB, closely matching its activity in RPMI+10%LB (Figure 2.2b). Similarly, replacement of glucose with maltose in RPMI+10%LB led to increased activity of the CcpA-1 iModulon. The change in glucose concentration, however, had little effect on the activity level of the CcpA-2 iModulon, suggesting that the CcpA-1 iModulon represents direct glucose-responsive CcpA activity, whereas the CcpA-2 iModulon may reflect its glucose-independent activity.

In addition to CcpA activity, we also confirmed the activity of the GR iModulon. The GR

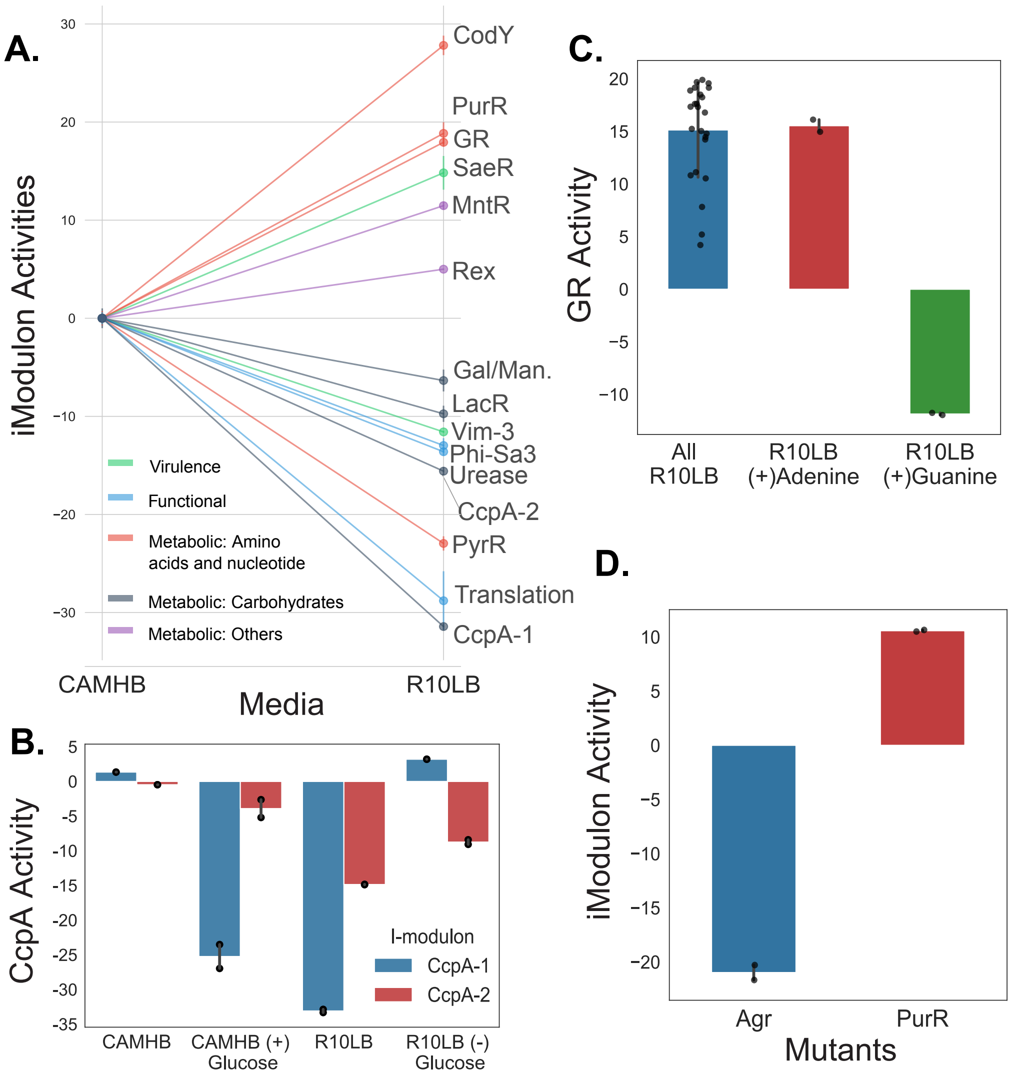


Figure 2.2: Differential activation of iModulons in different media conditions. (a) iModulons from LAC strain with statistically significant ($p-value < 0.05$) differential activation in CAMHB versus RPMI+10%LB. (b) Addition of glucose reduced the activity of CcpA-1 in CAMHB (blue bars). Conversely, replacing glucose with maltose led to higher CcpA-1 activity in RPMI+10%LB. CcpA-2 activity did not change in response to glucose concentration (red bars). (c) The bar plot shows the activity level of GR iModulon, which contains the genes under the control of guanine riboswitch (xpt and pbuG). Though many different conditions can affect the GR iModulon activity (blue bar), it sharply decreases when guanine is added to the media. Addition of adenine has no effect. Black dots in panel b and c represent values from individual samples and error bars represent standard deviation.(d) External validation of Agr and PurR iModulon activity in the respective *agr* and *purR* mutants.

iModulon contains genes involved in the purine salvage pathway (*xpT*, *pbuX*), peptide transport (*oppB*), and LAC specific virulence factor *ssl11*. The two genes in the salvage pathway have been previously demonstrated to be under the control of the guanine riboswitch in *S. aureus* strain NRS384[15]. The presence of this riboswitch was confirmed using the online RiboSwitch Finder (SI Appendix, Figure A.2c,d)[24]; no riboswitches were detected for the other two genes. The activity of the iModulon was attenuated by guanine supplementation (25ug/mL) while the addition of adenine had no effect, demonstrating a guanine-specific activity of the iModulon (Figure 2.2c).

We additionally validated activities of Agr and PurR iModulons using publically available expression profiling datasets (GSE18793 and GSE132179)[25, 26]. These datasets include expression profiles comparing wild type USA300 strains to their isogenic agr and purR mutants. As a form of external validation, we did not incorporate these data into the model. Instead we projected the expression data onto the model to convert the gene expression levels to iModulon activity levels (see Materials and Methods). Compared to their respective wild types, PurR iModulon had the largest increase in activity in purR::bursa strain and Agr iModulon showed the largest drop in activity in the strain with disrupted agr system, demonstrating that the model can capture activities of these iModulons in the conditions not included in the model (Figure 2.2d).

2.3.3 Integration of iModulons with genome-scale metabolic models reveal systems-level properties of metabolic regulation

Genome-scale metabolic models (GEMs) are knowledge-bases reconstructed from all known metabolic genes of an organism, systematically linking metabolites, reactions, and genes[27]. Integration of iModulons with these metabolic models allows us to probe the interaction between the regulatory and metabolic networks. To visualize this crosstalk at the systems level, we overlaid the iModulons onto central metabolism and amino acid metabolism pathways of the *S. aureus* metabolic reconstruction iYS854 (Figure 2.3a)[28]. The CcpA-1 and CodY iModulons dominate regulation of the genes in these metabolic subsystems of *S. aureus*. The two CcpA iModulons controlled many of the genes in carbon metabolism. The genes required for the tricarboxylic acid (TCA) cycle were found primarily in the CcpA-1 iModulon, with the exception of genes encoding fumarase and malate dehydrogenase. Additionally, the CcpA-1 iModulon contained genes required for degradation of gluconeogenic amino acids (serine, histidine, and alanine) and secondary metabolites (chorismate and N-acetyl-neuraminic acid). Also included were genes encoding two key gluconeogenic enzymes - phosphoenolpyruvate (PEP) carboxykinase and fructose-1,6-bisphosphatase. Genes involved in transport of alternate carbon sources were also present.

In contrast to catabolic CcpA-regulated genes, the iModulon associated with CodY regulation was dominated by genes participating in biosynthesis of amino acids lysine, threonine, methionine, cysteine, histidine, and branched chain amino acids (BCAA) isoleucine, leucine, and valine[13]. Regulation of interconversion between glutamine and glutamate (*gltA*), a key component of nitrogen balance and assimilation, was also a part of the CodY iModulon.

While the two iModulons (CcpA-1, CodY) did not share any genes, they intersected at some key metabolite nodes in central metabolism, including pyruvate, glutamate, histidine, and arginine. Genes in the CcpA-1 iModulon encode enzymes that generate pyruvate from amino acids and use the pyruvate to generate energy through fermentation, synthesize glucose via gluconeogenesis, or synthesize fatty acids via malonyl-coA. Enzymes encoded by genes in the

CodY iModulon, on the other hand, redirect pyruvate to instead synthesize BCAA (isoleucine, leucine and valine). Similarly, glutamate is directed towards the urea cycle by CcpA-1 and towards biosynthesis of the aspartate family amino acids by CodY. While genes required for catabolism of histidine are in the CcpA-1 iModulon, genes encoding histidine biosynthesis is instead part of the CodY iModulon.

2.3.4 Genome-scale metabolic models compute flux-balanced state that reflect regulatory actions of CcpA

Metabolic network reconstructions can be converted into genome-scale models that allow for the computation of phenotypic states[29]. We can compute the optimal flux through the metabolic network using flux-balance analysis (FBA)[28]. In particular, we can compute the metabolic state that is consistent with nutrient sources in a given environment to support optimal bacterial growth. In the previous CcpA iModulon validation experiment, we observed that changing the carbon source from glucose to maltose in RPMI+10%LB also led to an unexpected spike in activity of the iron-responsive Fur iModulon (SI Appendix, Figure A.3a). To investigate whether there was a possible metabolic role explaining the increase in Fur activity, we generated two condition-specific genome scale metabolic models (csGEMs) starting with iYS854[28]. For both csGEMs, we computed the state of the metabolic network that supports growth in RPMI+10%LB, with either glucose or maltose as the main carbon source (Methods and Materials). We assumed that CcpA-1 repression was active only when glucose was the main glycolytic nutrient source (and the corresponding set of reactions was shut off). Reaction fluxes across the network were then sampled using flux-balance analysis, assuming that the bacterial objective was biomass production[30]. Sampling accounts for different network flux distributions that can achieve the same optimal solutions (i.e., identical biomass production rates).

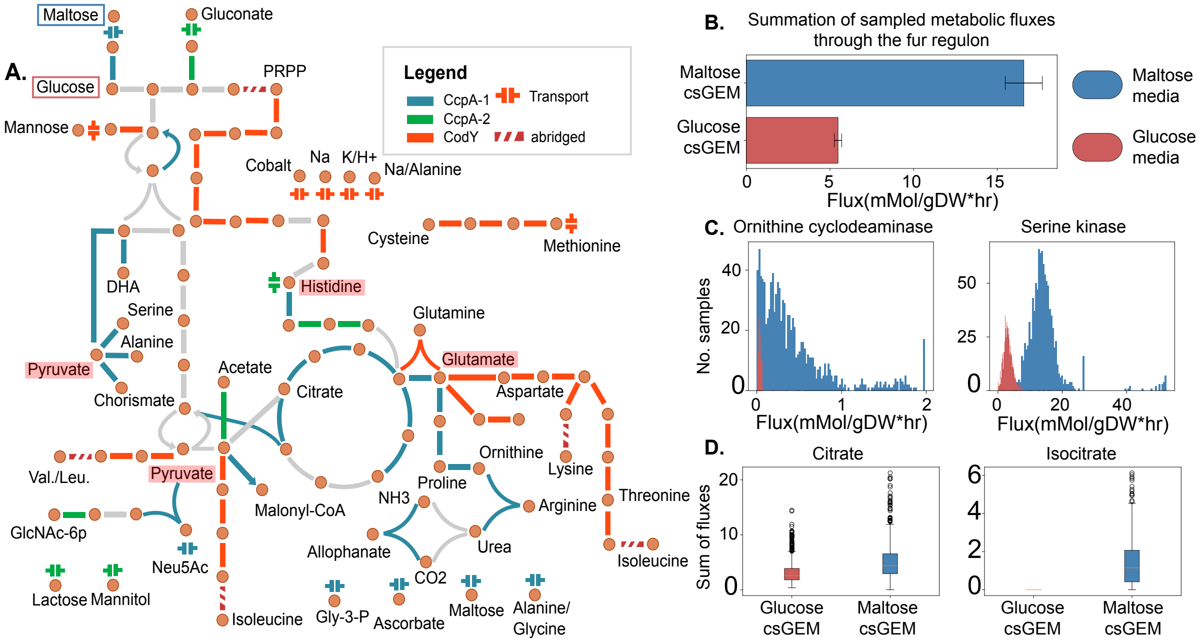


Figure 2.3: Regulation of central metabolism and its interaction with other metabolic subsystems. (a) Overlay of iModulons onto the map of central metabolism and amino acid metabolism of *S. aureus*. The two main regulators of these metabolic subsystems, CcpA (blue/green) and CodY(orange), control central carbon and nitrogen metabolism respectively. These two iModulons intersect at key metabolic nodes-pyruvate, histidine, and glutamate (highlighted in red). Entry points of sugars used in the next section, glucose and maltose, are highlighted with red and blue boxes respectively. (b) Activity of reactions associated with the Fur iModulon in presence of different carbon sources: maltose and glucose. The bars represent sum of median sampled fluxes through reactions catalyzed by enzymes in the Fur iModulon. Unexpected increase in Fur iModulon activity when carbon source was switched from glucose to maltose is recapitulated through metabolic modeling. (c) Reactions associated with the Fur iModulon with the largest increase in simulated flux in glucose media. (d) Calculated proxy for intracellular metabolite concentrations.

Under these conditions, the sum of sampled fluxes through reactions associated with the Fur iModulon was significantly higher in maltose media (Kolmogorov-Smirnov test, $p < 0.01$, $statistics > 0.9$), confirming that the spike in Fur activity could be a result of metabolic flux rewiring (Figure 2.3b). In particular, fluxes through serine kinase (*sbnI*, a precursor metabolic step of staphyloferrin B biosynthesis) and ornithine cyclodeaminase (*sbnB*) were significantly increased (Figure 2.3c). These changes came as a result of flux rewiring away from deactivated metabolic steps. For example, due to arginase (*rocF*) deactivation, the flux through half of the urea cycle and ornithine cyclodeaminase was lower. Similarly, serine deaminase (*sdaB*) - located two metabolic steps downstream of serine kinase - was deactivated due to simulated down-regulation of genes in the CcpA-1 iModulon, and flux through phosphoglycerate dehydrogenase, serine kinase, and phosphoserine phosphatase was decreased. We computed the sum of fluxes producing each metabolite as a proxy for intracellular concentrations and found that the calculated values were significantly larger in maltose media for 68 metabolites including ammonium, glutamate, and isocitrate. The majority of the TCA cycle was shut off in the glucose-specific GEM (due to simulated repression of *citB*, *icd*, *odhA*, *sdhABCD*, and *sucCD*), and therefore the concentration proxy for isocitrate was essentially null, while that of citrate was not (Figure 2.3d). Previous studies have shown that *citB* deletion results in increased intracellular concentration of citrate [31]. Apart from being an intermediate in the TCA cycle, citrate can be utilized in the model as a precursor to staphyloferrin A and staphyloferrin B biosynthesis (which are included in the Fur iModulon), or it can be converted back to oxaloacetate and acetate via citrate lyase. All three routes were part of the solution space, with citrate lyase carrying the largest median flux. Taken together, these modeling simulations suggest that utilizing maltose instead of glucose induces metabolic flux rewiring towards reactions associated with the Fur iModulon.

2.3.5 An iModulon details possible scope and functions of sigma factor σS

Global stress response in *S. aureus* is modulated by the alternate sigma factor σB [32, 33]. Though two other sigma factors, σS and σH, have been recognized in this organism, their exact functions and full regulon are not as well understood[34, 35]. We identified two iModulons that correspond to sigma factors σ B and σS. The SigB iModulon contained genes encoding σB

($sigB$), anti-σ B ($rsbW$), and anti-σ B antagonist ($rsbV$). The activity of SigB iModulon was correlated with *sigB* expression ($PearsonR = 0.55$, $p - value = 8.2e - 11$) (Figure 2.4a), with the highest activation in stationary phase ($OD600 = 1$). Furthermore, a conserved 29 bp motif was enriched from 28 unique regulatory regions of SigB iModulon genes (SI Appendix, Figure A.4a) (see Methods and Materials). As the regulatory role of σ B has been previously explored in detail[33, 36–39], we focused here on the less understood regulatory role of σS. Though σS is important for both intracellular and extracellular stress response, its full regulon has yet to be defined[35, 40]. ICA identified a large iModulon with 137 genes including *sigS* itself (which encodes σS). As with the SigB iModulon, expression of the *sigS* gene correlated to activity of the ICA-derived SigS iModulon (Pearson $R = 0.77$, $p - value = 4.26e - 22$) (Figure 2.4b). Previous studies have shown that CymR represses *sigS* expression and therefore may lead to its decreased activity[41]. We confirmed this relationship as the SigS iModulon activity was anti-correlated with the CymR iModulon activity ($PearsonR = -0.68$, $p - value = 8.23e - 10$) (SI Appendix, Figure A.4b).

To further characterize σS, we looked for conserved motifs in the regulatory regions of the genes in the iModulon and found a purine-rich 21 base-pair purine rich motif ($E-value = 7.7e-8$) in the regulatory region of at least 56 genes in the SigS iModulon (Figure 2.4c). Comparisons against a known prokaryotic motif database revealed that the *S. aureus* σS motif was most similar to that of the σ B (MX000071) motif in *B. subtilis* ($E - value = 1.62e - 02$) (see Materials and Methods). Next, we analyzed the distance between the center of the motif and the transcription initiation site. For most genes, the motif was present at or around 35 base-pairs upstream of the translation start site, though motifs were also found further upstream (Figure 2.4d).

Of the 137 genes in the iModulon, only 56 (41%) had an assigned function in the reference

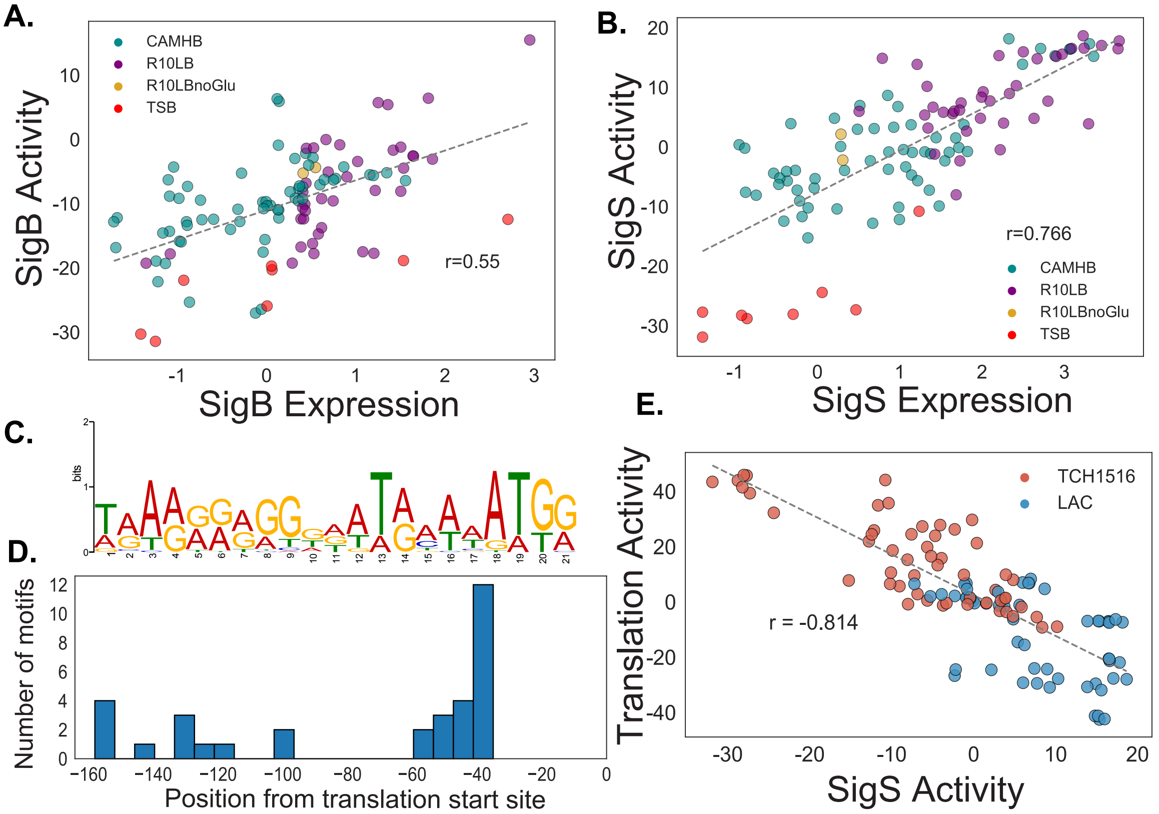


Figure 2.4: Profiling alternate sigma factor S. The expression levels of *sigB* (a) and *sigS* (b) genes and the activity levels of their respective iModulons show strong positive correlation. (c) The regulatory region (150 bp upstream of the first gene in operon) of genes in the SigS iModulon contained a conserved purine rich motif. (d) The positions (relative to transcription start-site) of the enriched motif within the regulatory sites of genes in the SigS iModulon. For many genes in the SigS iModulon, the motif was present 35 bp upstream of the translation start site. (e) 'Greed vs. Fear' trade-off is reflected in the activity of the Translation (greed) and SigS (fear) iModulons. LAC showed increased propensity for fearful bet-hedging strategy while TCH1516 relied on a more greedy strategy.

genome, further highlighting our limited understanding of σS functionality. However, many of the annotated gene products were key factors in controlling cellular state. These included factors regulating virulence (*sarA*, *sarR*, *sarX*), antimicrobial resistance (*cadC*, *blaI*), metabolism (*arcR*, *argR*), cell wall biogenesis (*vraRST*), biofilm formation (*icaR*), and DNA damage repair (*recX*). Genes encoding proteins critical for stress response such as universal stress protein (Usp), toxin MazF, competence proteins ComGFK, and cell division protein were also present.

The SigS iModulon also plays a critical role in the so-called 'fear vs. greed' trade-off in *S. aureus*. Previously described in *E. coli*, this trade-off describes the allocation of resources towards optimal growth (greed) versus allocation towards bet-hedging strategies to mitigate the effect of stressors in the environment (fear)[7, 42]. This balance is reflected in the transcriptome composition as an inverse correlation between the activities of the stress-responsive SigS iModulon and the Translation iModulon (Figure 2.4e). Unlike *E. coli*, however, this relationship was independent of growth rate, as growth rate had weak correlation with Translation iModulon expression activity ($PearsonR = 0.094$, $p-value = 0.514$). Interestingly, mapping this trade-off highlighted a possible difference in survival strategy between the two USA300 strains. TCH1516 tended towards a greedy strategy with high Translation iModulon activity while LAC was more likely to rely on bet-hedging, or fear.

2.3.6 ICA reveals organization of virulence factor expression

ICA captured systematic expression changes of several genes encoding virulence factors. Previous studies described over half a dozen transcription factors with direct or indirect roles in regulation of virulence factor expression in *S. aureus*[43]. The number of regulators, and their complex network of interactions, make it extremely difficult to understand how these genes are regulated at a genome scale. In contrast, ICA identified only three iModulons - named Agr, SaeR, and Vim-3 - that were mostly composed of virulence genes (Figure 2.5a). The activity level of Agr had extremely low correlation with that of SaeR and Vim-3, suggesting that Agr may have only limited cross-talk with the other two iModulons in our conditions (SI Appendix, Figure A.5a). However, the activity levels of SaeR and Vim-3 were negatively correlated (Pearson R = -0.57, p-value = 8.6e-11). As the two iModulons contain different sets of virulence factors, the negative correlation points to a shift in the virulence state where *S. aureus* may adopt different strategies to thwart the immune system. Collectively, the three virulence iModulons revealed coordinated regulation of 65 genes across the genome. These results suggest that the complexity behind virulence regulation can be decomposed into discrete signals and the virulence state of *S. aureus* can be defined as a linear combination of these signals.

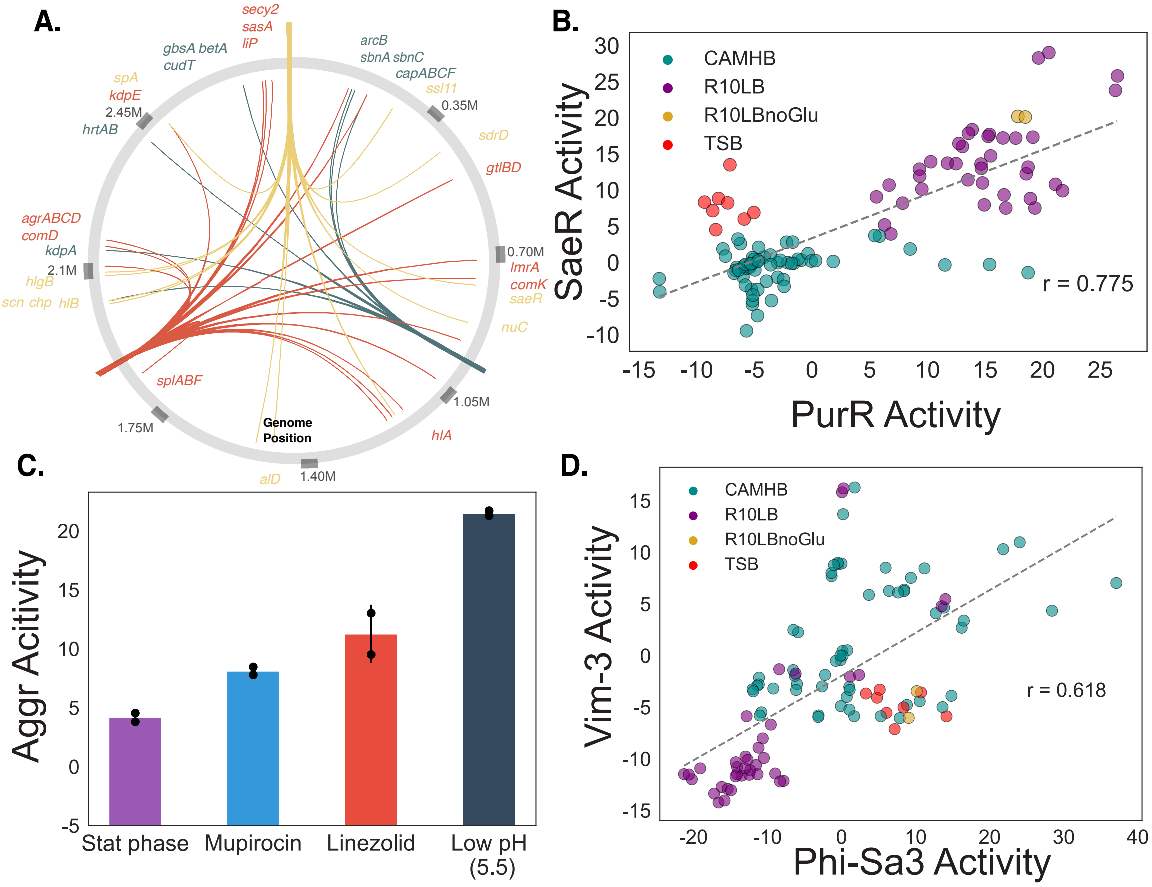


Figure 2.5: Global regulation of virulence factors. (a) The three virulence iModulons (SaeR, Agr, Vim-3) and the genomic positions of the genes in their respective iModulons are mapped. The signals encode over 25 virulence factor associated genes. (b) PurR iModulon activity is highly correlated with virulence iModulon SaeR. (c) Challenge with low pH, linezolid, and mupirocin leads to strong activation of agr in exponential growth phase. Interestingly, this activation is stronger than that induced by stationary phase ($O.D.600 = 1.0$). Activation of agr was much weaker under all other experimental conditions considered (top bar chart). (d) Co-activation of Phi-Sa3 iModulon with virulence iModulon Vim-3.

The SaeR iModulon contained 27 genes, including the genes for the SaeRS two-component system (TCS). The activity level of this iModulon strongly correlated with the expression level of *saeRS*, further supporting the idea that the genes in this iModulon are regulated (directly or indirectly) by SaeRS ($PearsonR = 0.80$, $p{-}value = 1.38e{-}25$). Furthermore, the virulence genes *chp*, *coa*, *ssl11*, *sbi*, *map*, *lukA*, and *scn*, previously reported to be under the control of SaeRS[44], were also found in this iModulon. The activity of SaeR iModulon was strongly associated with purine metabolism. PurR, the transcription factor that regulates the genes of purine biosynthesis, has been recently implicated in regulation of virulence factors[25, 45]. Consistent with this observation, the activity level of the SaeR iModulon correlated well ($PearsonR = 0.77$, $p - value = 8.9e - 23$) with the activity of the PurR iModulon (Figure 2.5b). Thus, SaeR may act as a bridge between virulence and metabolism.

Similarly, the Agr iModulon contained the *agrABCD* genes involved in regulation of the quorum sensing *agr* regulon[46, 47]. As most of our samples were collected during early- to mid-exponential growth phase, the Agr iModulon remained inactive in these conditions (SI Appendix, Figure A.5b). Only acidic conditions (pH 5.5) and treatment with translation inhibitors linezolid and mupirocin activated Agr during exponential growth (Figure 2.5c). Both pH- and translation inhibition-dependence of *agr* expression have been previously reported [48–51]. Unexpectedly, the Agr iModulon was activated to a much greater extent by these factors than high cell density (O.D. 1.0), for which its role in quorum sensing is extensively characterized.

The Vim-3 virulence iModulon consisted of genes required for siderophore and heme utilization (*sbnABC*, *hrtAB*), capsule biosynthesis (*cap8a*, *capBC*, *cap5F*) , and osmotic tolerance (*kdpA*, *betAT*, *gbsA*). The Vim-3 iModulon had maximal activity under hyperosmotic condition introduced by 4% NaCl and when grown to stationary phase (OD 1.0) in CAMHB (SI Appendix, Figure A.5c). The increased expression of capsule biosynthesis genes have shown to be responsive to change in osmotic pressure as well as iron starvation, which is consistent with the inclusion of iron scavenging and osmotic tolerance genes in the iModulon with the capsular biosynthesis genes[52, 53].

We further identified a prophage Phi-Sa3 associated iModulon as a new putative iModulon required for virulence. The Phi-Sa3 iModulon consists of genes in the Phi-Sa3 prophage and several genes encoding DNA replication and repair enzymes. Excluded from the iModulon were the virulence factors that were horizontally acquired along with the phage (scn and chp)[54], which now fell under the control of SaeR. Of the four phages in *S. aureus* strain Newman, Phi-Sa3 is the only prophage that is unable to generate complete viral particles when challenged with DNA damaging agent mitomycin[55]. However, evidence suggests that this prophage is still active in USA300 strains and its genes are expressed during lung infection, where it may play a role in establishing virulence[56]. Corroborating this hypothesis, we found that the activity of the Phi-Sa3 iModulon correlated highly with the Vim-3 iModulon ($PearsonR = 0.62$, $p - value = 9.9e - 13$)(Fig. 5e). As the Phi-Sa3 iModulon does not contain any virulence genes, the phage itself may play an accessory role in establishing virulence.

2.3.7 ICA model provides a platform for *in-vivo* data interpretation

Transcriptomic models based on ICA can also be used to interpret new *in-vivo* and *ex-vivo* expression profiles, leading to greater clarity when compared to analysis with graph based TRN model (Appendix A.2). Expression profiling data can be projected onto the iModulon structure of the TRN, derived from our dataset, to convert the values from gene expression levels to iModulon activity levels (see Materials and Methods). This projection can supplement gene differential expression analysis by identifying regulators that are driving the large changes in gene expression often seen *in-vivo*.

We projected a microarray data (GSE61669) taken at 24 hour post-infection from a rabbit skin infection model model[57]. After 24 hours, 1232 differentially expressed genes were reported. Projection of the data on to the model showed that these changes in differential expression are being driven by simultaneous activation of CodY and Fur iModulons and inactivation of SigB, PurR Agr and Translation iModulons (Figure 2.6a).

In time-course data, projecting expression data onto the model can also help us understand the dynamics of different regulators during infection. We projected previously published time course microarray data collected from *S. aureus* USA300 LAC grown in Tryptic Soy Broth (TSB), human blood, and serum[58]. Bacteria grown to an exponential phase in TSB was used as inoculum for all samples; we used this as our new base condition for the projected data. Therefore, all iModulon activity levels in this set represent log2 fold change in activity from this base condition. Once transferred to serum, the activities of Fur and CodY iModulons in serum increased dramatically with Fur being activated immediately after exposure to serum while CodY activating slowly over time to reach a similar level as Fur by 2 hours (Figure 2.6b). The large change in activity coupled with the sizeable number of genes in each iModulon (80 and 45 genes in CodY and Fur, respectively) indicates that *S. aureus* reallocates a considerable portion of its transcriptome to reprogram amino acid and iron metabolism in serum. PurR and SaeR activity also increased, though their magnitude of change was dwarfed by the changes in activity of CodY and Fur. Agr activity, on the other hand, declined and remained low over the two hour period. Because agr positively regulates a number of virulence genes, dynamic changes in its activity level could be expected in serum. However, consistent with the model prediction, previous stud-

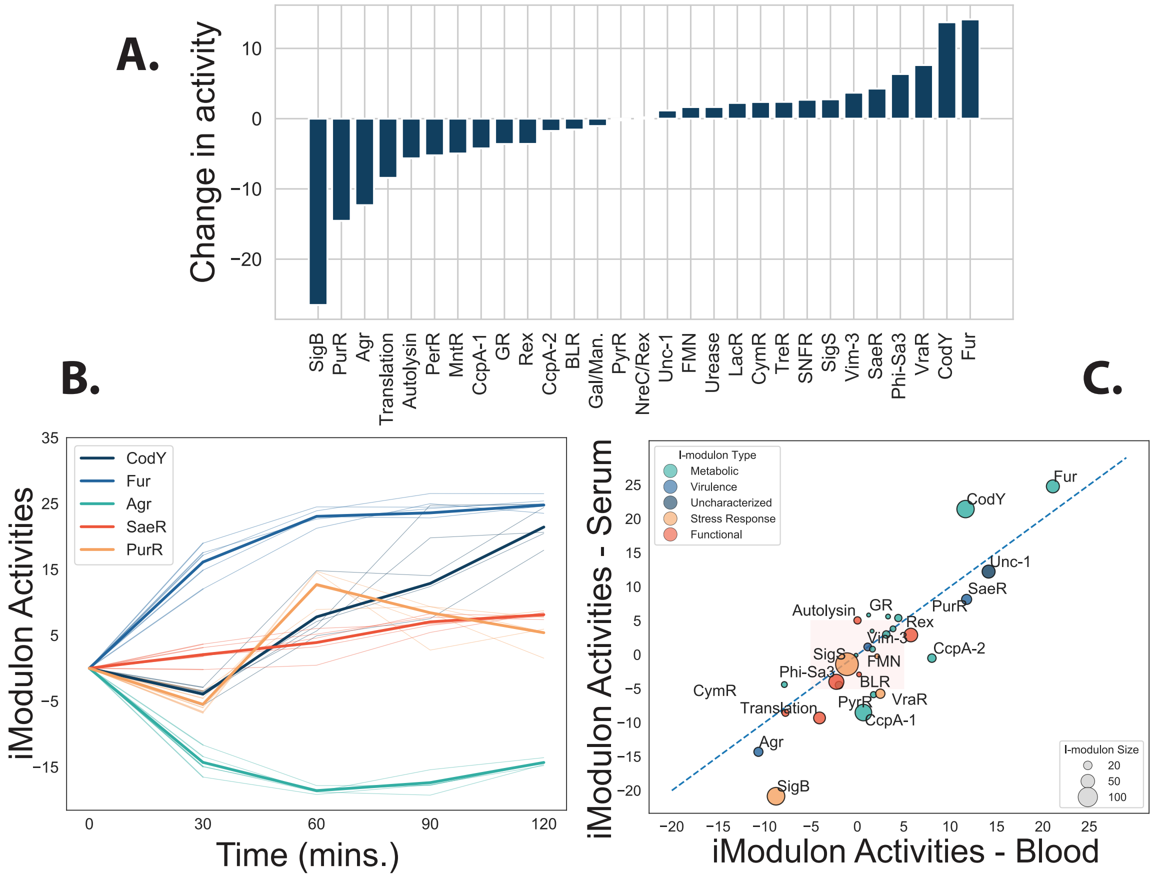


Figure 2.6: ICA analysis of *in vivo* and *ex vivo* data. (a) Change in iModulon activities at 24 hours post infection in a rabbit skin infection model. (b) Activity levels of select iModulons in serum over the two hour time period. The thick line represents the mean activity across all replicates and the thin line represents activity in each individual replicate (n=4). Activity levels were around the inoculum values, (c) Comparison of iModulon activity between serum and blood and two hour time point. The dashed red line is the 45 degree line; iModulons below the line have higher activity in blood and those above the line have higher activity in serum. Red shaded area contains iModulons with less than 5 fold change in activity in both conditions.

ies have demonstrated that agr transcription is dampened in human serum due to sequestration of auto-inducing peptide (AIP) by human apolipoprotein B[59, 60].

We next calculated the differences in iModulon activities in blood and serum at the final two hour time-point. Fur, CodY, PurR, SaeR, and Agr had similar activity levels in both blood and serum (Figure 2.6b). Therefore, the activity of these regulators are likely governed by the non-cellular fraction of the blood. iModulons PyrR, SigB, Translation, VraR, CcpA-1, and CcpA-2 had higher activity levels in blood than in the serum (Figure 2.6c). Glucose concentration in blood is lower than in serum, which likely explains the shift in CcpA-1 activity[61]. The lower glucose concentrations relieves CcpA mediated repression of its regulon, leading to higher expression. The shift in the PyrR iModulon also corroborates previous studies, which demonstrated that *S. aureus* strain JE2 (a derivative of LAC) requires more pyrimidine when growing in blood than in serum[62]. The signals or cues driving the change in activity of the other iModulons (SigB, Translation, VraR, and CcpA-2) remains unknown.

Overall, the imodulon analysis revealed that during acute infection, CodY and Fur play key roles in rewiring the *S. aureus* metabolism in serum and blood when compared to TSB, while SaeR (and not Agr) drives the virulence gene expression. In addition, SigB, Translation, and VraR iModulons are uniquely activated by the cellular fraction of the blood and may thus be responding to unique stresses they impart.These observations however are limited as the baseline for comparisons for most of these analyses were *in vitro* growth in TSB. Though the differentially activated iModulons may point to important roles that each of the associated regulators play during acute infection, further analysis is still required to understand their relative contribution. The model is also limited in that it is currently blind to the regulators that are not captured in any of the 29 iModulons. This limitation will be alleviated over time as we incorporate more sequencing data that is being generated at an ever increasing pace.

2.4 Discussion

Here, we described an ICA-based method to elucidate the organization of the modules in TRN in *S. aureus* USA300 strains. Using this method, we identified 29 independently modulated sets of genes ('iModulons') and their activities across the sampled conditions. This framework for exploring the TRN provides three key advantages over traditional methods, especially when working with non-model organisms: (1) the method provides an explanatory reconstruction of the TRN; (2) it is an untargeted, and therefore unbiased, approach; and (3) the approach utilizes expression profiling data, an increasingly ubiquitous resource.

First, iModulons quantitatively capture the complexities of transcriptional regulation and enable a new way to systematically query the transcriptome. By recasting the data in terms of explanatory iModulons, we gained a deeper understanding of large changes in transcription profiles between CAMHB bacteriologic media and the more physiologically relevant mammalian tissue culture-based media RPMI+10%LB. The analysis reduced the number of features needed to capture most of the information in the transcriptome from hundreds of genes to fifteen iModulons. Additionally, quantified activity levels of iModulons also enabled integration of regulatory activity with metabolic models and revealed coordination between metabolic and regulatory networks. Such reduction in complexity and the integration of different aspects of *S. aureus* biology (e.g., virulence, metabolism, stress response, etc.) will be crucial to understanding the mechanisms that enable successful infection *in-vivo*.

Second, this method presents a platform for untargeted, global analysis of the TRN. Due to its untargeted nature, we also identified two key virulence features of *S. aureus.* ICA revealed coordinated regulation of genes in capsule biosynthesis, osmotic tolerance, and iron starvation (Vim-3 iModulon). Both capsule formation and siderophore scavenging are important in nasal colonization[63, 64]. Similarly, growth of *S. aureus* in Synthetic Nasal Medium (SNM3) increases the expression of genes required for osmotolerance. Therefore, the Vim-3 iModulon may represent a concerted regulation of genes required for successful nasal colonization. We also identified the Phi-Sa3 phage iModulon, whose activity level correlated with that of the Vim-3 iModulon. The

Phi-Sa3 iModulon did not include the virulence genes (e.g. *sak*, *scn*, etc.) that were acquired with the phage, suggesting that phage replication genes were expressed independently of the virulence genes. Given that its activity was correlated with Vim-3, this phage may also play an important role in nasal colonization.

Lastly, ICA uses RNA-seq data to extract information about the TRN, making it more accessible to non-model organisms including *S. aureus*. Reconstructing the TRN with traditional methods is highly resource intensive, as they require targeted antibodies or specialized libraries of plasmids containing all transcription factors of interest[65].. While these approaches have given us great insights into TRNs of model organisms like *E. coli*[10], such comprehensive data is not available for most microbes. Several studies have attempted to circumvent this by comparing the expression profiles of wild-type *S. aureus* strains with their counterpart through transcription factor knockout or introduction of a constitutively active transcription factor. However, these approaches often overestimate the regulatory reach of the transcription factor, as such genetic changes can trigger the differential expression of genes not directly under the regulator's control. By identifying iModulons consisting of independently regulated sets of genes, ICA-based method improves on these approaches as it able to segregate specific regulator targets[7]. While a large number of expression profiles are required to build such a model, a rapidly growing number of expression profiles are already publicly available on Gene Expression Omnibus (GEO). Indeed, utilizing only RNA-seq data, we predicted the previously unknown regulon of stress-associated sigma factor σS and its possible roles in biofilm formation, growth rate control, and general stress response. With the growing number of available expression profiles, such characterizations can be extended to other undefined or poorly defined regulons. Therefore, in the absence of a comprehensive set of targeted antibodies against *S. aureus* transcription factors, reanalyzing the publicly available database with ICA could be used to further reconstruct its TRN.

We have shown that ICA based decomposition can be utilized to build a quantitative and explanatory model of *S. aureus* TRN from RNA sequencing data. Application of this model enabled us to query metabolic and regulatory crosstalk, discover new potential regulons, find coordination between metabolism and virulence, and unravel the *S. aureus* response to during growth in blood. Due to this versatility, this model and other models generated through this framework, may prove to be a powerful tool in any future studies of *S. aureus* and other non-model organisms.

2.5 Materials and Methods

2.5.1 RNA extraction and library preparation

S. aureus USA300 isolates LAC, TCH1516 and ALE derivatives of TCH1516 (SNFR and SNFM) were used for this study. The growth conditions and RNA preparation methods for data acquired from Choe et al. has been previously described[66]. Detailed growth conditions, RNA extraction and library preparation methods for other samples have also been already described[67]. Briefly, an overnight culture of *S. aureus* was used to inoculate a pre culture and were grown to mid-exponential growth phase ($OD600 = 0.4$) in respective media (CAMHB, RPMI + 10% LB, or TSB). Once in mid-exponential phase, the preculture was used to inoculate the media containing appropriate supplementation or perturbations. Samples were collected at O.Ds and time-points indicated in the metadata. All samples were collected in biological duplicates originating from different overnight cultures. Sample for control conditions were collected for each set to account for batch effect.

2.5.2 Determining Core Genome with Bi-Directional BLAST Hits (BBH)

To combine the data from the two strains, core genome containing conserved genes between the LAC (GenBank: CP035369.1 and CP035370.1) and TCH1516 (GenBank: NC_010079.1, NC_012417.1, and NC_010063.1) were first established using BBH[68]. In this analysis, all protein sequences of CDS from both genomes are BLASTed against each other twice with each genome acting as reference once. In this analysis, all protein sequences of CDS from both genomes are BLASTed against each other twice with each genome acting as reference once. Two genes were considered conserved (and therefore part of the core genome) if (1) the two genes have the highest alignment percent to each other than to any other genes in the genome, and (2) the coverage is at least 80%.

2.5.3 RNA Sequencing Data Processing

RNA sequencing pipeline used to analyze and perform QC/QA has been described in detail previously[67]. Briefly, the sequences were aligned to respective genomes, LAC or TCH1516 using Bowtie2[69, 70]. The samples from ALE derivatives, SNFM and SNFR, were aligned to TCH1516. The aligned sequences were assigned to open reading frames using HTSeq-counts [71]. Differential expression analysis was performed using DESeq2 with p-value threshold of 0.05 and an absolute fold change threshold of 2[72]. To create the final counts matrix, counts from conserved genes in LAC samples were represented by the corresponding ortholog in TCH1516. The counts for accessory genes were filled with 0s if the genes were not present in the strain (i.e. LAC specific genes had counts of 0 in TCH1516 samples and vice versa). Finally, to reduce the effect of noise, genes with average counts per sample less than 10 were removed. The final counts matrix with 2581 genes was used to calculate Transcripts Per Million (TPM).

2.5.4 Computing robust components with ICA

Procedure for computing robust components with ICA has been described in detail previously [7]. Log2(TPM + 1) values were centered to strain specific reference conditions and used as input of ICA decomposition. These conditions are labeled: 'USA300_TCH1516_CAMHB_U01-Set000_Control_1', 'USA300_TCH1516_CAMHB_U01-Set000_Control_2' for TCH1516 and 'USA300_LAC_CAMHB_U01-Set001_Control_1', 'USA300_LAC_CAMHB_U01-Set001_Control_2' for LAC. Next, Scikit-learn (v0.19.0) implementation of FastICA algorithm was used to calculate independent components with 100 iterations, convergence tolerance of 10-7, log(cosh(x)) as contrast function, and parallel search algorithm[73, 74]. The number of calculated components were set to the number of components that reconstruct 99% of variance as calculated by principal component analysis. The resulting S-matrices containing source components from the 100 iterations were clustered with Scikit-learn implementation of DBSCAN algorithm with epsilon of 0.1, and minimum cluster seed size of 50 samples (50% of the number of random restarts). If necessary, the component in each cluster was inverted such that the gene with the maximum absolute weighting the component was positive. Centroids for each cluster was used to define the final weightings for S and corresponding A matrix. The whole process was repeated 100 times to ensure that the final calculated components were robust. Finally, components with activity levels that deviated more than 5 times between samples in the same conditions were also filtered out.

2.5.5 Determining independently modulated sets of genes

ICA enriches components that maximize the non-gaussianity of the data distribution. While most genes have weightings near 0 and fall under gaussian distribution in each component, there exists a set of genes whose weightings in that component deviates from this significantly. To enrich these genes, we used Scikit-learn's implementation of the D'Agostino K2 test, which measures the skew and kurtosis of the sample distribution[75]. We first sort the genes by the absolute value of their weightings and perform the K2 test after removing the gene with the highest weighting. This was done iteratively, removing one gene at a time, until the K2 statistic falls below a cutoff. We calculated this cutoff based on sensitivity analysis on agreement between enriched iModulon genes and regulons inferred by RegPrecise[76]. For a range of cutoff (between 200-600), we ran the iterative D'Agostino K2 test on all components and checked for statistically significant overlap of iModulons with the regulons predicted by RegPrecise using Fisher's Exact Test. For iModulons with significant overlap, we also calculated precision and recall. The cutoff of 280 which led to the highest harmonic average between precision and recall (F1-score) was chosen as the final cutoff.

2.5.6 Designating biological annotations to iModulons

To designate proper annotations to iModulons, we first compiled a dataset containing previously predicted features such as regulons, genomic islands and plasmids. The regulons in the datasets were inferred by either RegPrecise algorithm and by RNA-seq analysis of transcription factor knockout strains or strains with constitutively active transcription factors[66, 77–80]. Genomic islands were determined by online IslandViewer4 tool[81] and phages were identified with PHASTER[82]. For studies using different strains of *S. aureus* orthologs for TCH1516 and LAC were determined using BBH. The enriched genes in iModulons were compared against this dataset for significant overlap using Fisher's Exact Test with FDR of 10-5. With this analysis 15 iModulons were enriched with high confidence ($precision >= 0.5$, $recall >= 0.2$) and 7 were enriched with low confidence. Additionally, iModulons containing genes with shared functions (e.g. Translation and B-lactam Resistance) were annotated manually.

2.5.7 Differential Activation Analysis

Distribution of differences in iModulon activities between biological replicates were first calculated and a log-norm distribution was fit to the differences. In order to test statistical significance, absolute value of difference in activity level of each iModulon between the two samples were calculated. This difference in activity was compared to the log-normal distribution from above to get a p-value. Because differences and p-value for all iModulons were calculated, the p-value was further adjusted with Benjamini-Hochberg correction to account for multiple hypothesis testing problem. Only iModulons with change in activity levels greater than 5 were considered significant.

2.5.8 Motif Enrichment and Comparison

Genes were first assigned to operons based on operonDB[83, 84]. For iModulon specific motif enrichments, 150 base pairs segment upstream of all the genes in the iModulons were collected. To avoid enriching ribosome binding sites, the segment started from 15 base pairs upstream of the translation start site. For genes in minus strand, the reverse complement of the sequence was used instead. If genes were part of an operon, then only the segment in front of the first gene in the operon was used. Motifs and their positions were enriched from these segments using the online Multiple Em for Motif Elicitation (MEME) algorithm[85, 86]. The following default parameters were used: -dna -oc -mod zoops -nmotifs 3 -minw 6 -maxw 50 -objfun classic -revcomp -markov_order 0. Enriched motifs were compared to combined prokaryotic databases-

CollecTF, Prodoric (release 8.9), and RegTransBase (v4) using TomTom[87–90]. The parameters for TomTom were as follows: -oc -min-overlap 5 -mi 1 -dist pearson -evalue -thresh 10.0.

2.5.9 Metabolic modeling

We modeled growth in RPMI supplemented with iron, manganese, zinc, and molybdate by setting the lower bound to the corresponding nutrient exchanges in iYS854 to -1 mmol/gDW/hr (the negative sign is a modeling convention to allow for the influx of nutrients) [28], and -13 mmol/gDW/hr for oxygen exchange (as measured experimentally). Additionally, to account for the utilization of heme by *S. aureus* terminal oxidases, we removed heme A from the biomass reaction and added as a reactant in the cytochrome oxidase reaction with the stoichiometric coefficient obtained from the biomass reaction[91]. Next, we constructed two condition-specific GEMs (csGEMs) to compare two conditions with: 1) D-glucose as the main glycolytic source and; 2) maltose as an alternative carbon source. In the first condition, we set the lower bound to D-glucose exchange to -50 mmol/gDW/hr. Assuming that in the presence of D-glucose, ccpA mediates the repression of multiple genes [22, 92], we set the upper and lower bounds of the reactions encoded by genes of the ccpA iModulon to 0. Specifically, we only turned off the set of 44 reactions obtained by running the "cobra.manipulation.find_gene_knockout_reactions()" command from the cobrapy package [93], feeding it the model and the 52 modeled genes which form part of the ccpA iModulon. As such, we implemented a method similar to the switch-based approach [94, 95], in which the boolean encoding for the gene-reaction-rule is taken into account (i.e. isozymes, and protein complexes). Shutting down all of the reactions yielded a model which could not simulate growth. We thus gap-filled the first csGEM with one reaction (AcCoa carboxylase, involved in straight chain fatty acid biosynthesis). To simulate the second condition in which maltose serves as the main glycolytic source, we set the lower bound of maltose exchange to -50 mmol/gDW/hr and blocked D-glucose uptake. No regulatory constraints were added. Flux-balance analysis was implemented with the biomass formation set as the functional network objective, and fluxes were sampled in both csGEMs 1000 times using the "cobra.sampling.sample" command. To normalize flux values across conditions, we divided all fluxes by the simulated growth rate. We compared the flux distribution of each reaction in the two csGEMs using the Kolmogorov-Smirnov nonparametric test, yielding 93 reactions with significantly differing flux distributions ($p - value < 0.001$) having a statistic larger than 0.99. To identify whether there is a metabolic basis for the difference the Fur iModulon stimulation between conditions, we identified a set of 34 reactions encoded by the 41 modeled genes which are part of the Fur iModulon (again using the switch-based approach).

2.5.10 Targeted High-Performance Liquid Chromatography (HPLC)

For glucose detection, samples were collected every 30 minutes and filtered as described above. Growth media was syringe-filtered through 0.22 µm disc filters (Millex-GV, Millipore-Sigma) to remove cells. The filtered samples were loaded onto a 1260 Infinity series (Agilent Technologies) high-performance liquid chromatography (HPLC) system with an Aminex HPX-87H column (Bio-Rad Laboratories) and a refractive index detector. The system was operated using ChemStation software. The HPLC was run with a single mobile phase composed of HPLC grade water buffered with 5 mM sulfuric acid (H2SO4). The flow rate was held at 0.5 mL/minute, the sample injection volume was 10 uL, and the column temperature was maintained at 45°C. The identities of compounds were determined by comparing retention time to standard curves of glucose. The peak area integration and resulting chromatograms were generated within Chem-

Station and compared to that of the standard curves in order to determine the concentration of each compound in the samples.

2.5.11 Microarray data analysis and projection

All microarray data was downloaded from GEO repository (GSE25454, GSE61669, and GSE18793) and processed with the Affy package in R to get gene expression level[58, 96]. GSE25454 dataset consists of microarray data from samples grown to exponential phase in TSB (TSB 0hr) and transferred to either blood, serum or TSB. Samples were then collected every 30 mins for 2 hours. The data was centered on 'TSB 0 hr' time-point. GSE61669 data consists of expression profile from 24 hour rabbit skin infection. This data was centered on the expression profile from the inoculum. Lastly, GSE18793 expression profile consists of data comparing WT LAC and its isogenic agr mutant. This data was centered around the WT expression profile. Data projection was used to convert centered gene expression values to iModulon activity level as described before[7].

2.5.12 Data and Code Availability

All RNA-seq data have been deposited to the Short Read Archive (SRA). All RNA-seq data were deposited to Sequence Read Archive (SRA). Custom code of ICA analysis can be found on github (https://github.com/SBRG/precise-db).

2.6 Acknowledgements

Chapter 2, in part, is a reprint of material published in: **Saugat Poudel**, Hannah Tsunemoto, Yara Seif, Anand Sastry, Richard Szubin, Sibei Xu, Henrique Machado, Connor

Olson, Amitesh Anand, Joe Pogliano, Victor Nizet, and Bernhard O. Palsson. "Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response." *Proc. Natl. Acad. Sci. U. S. A.* 117, 17228–17239 (2020). The dissertation author was the primary author.

2.7 References

1. Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L. & Fowler, V. G. Staphylococcus aureus infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* **28,** 603–661 (July 2015).

2. Krismer, B., Weidenmaier, C., Zipperer, A. & Peschel, A. The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. en. *Nat. Rev. Microbiol.* **15,** 675–687 (Oct. 2017).

3. Dastgheyb, S. S. & Otto, M. Staphylococcal adaptation to diverse physiologic niches: an overview of transcriptomic and phenotypic changes in different biological environments. en. *Future Microbiol.* **10,** 1981–1995 (Nov. 2015).

4. Goerke, C. & Wolz, C. Adaptation of Staphylococcus aureus to the cystic fibrosis lung. en. *Int. J. Med. Microbiol.* **300,** 520–525 (Dec. 2010).

5. Burian, M., Wolz, C. & Goerke, C. Regulatory adaptation of Staphylococcus aureus during nasal colonization of humans. en. *PLoS One* **5,** e10040 (Apr. 2010).

6. Ibarra, J. A., Pérez-Rueda, E., Carroll, R. K. & Shaw, L. N. Global analysis of transcriptional regulators in Staphylococcus aureus. en. *BMC Genomics* **14,** 126 (Feb. 2013).

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

8. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. en. *Nat. Commun.* **9,** 1090 (Mar. 2018).

9. Anand, A., Chen, K., Yang, L., Sastry, A. V., Olson, C. A., Poudel, S., Seif, Y., Hefner, Y., Phaneuf, P. V., Xu, S., Szubin, R., Feist, A. M. & Palsson, B. O. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. en. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 25287–25292 (Dec. 2019).

10. Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., Garcıa-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P., Ishida-Gutiérrez, C., Velázquez-Ramırez, D. A., Del Moral-Chávez, V., Bonavides-Martınez, C., Méndez-Cruz, C.-F., Galagan, J. & Collado-Vides, J. RegulonDB

v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. en. *Nucleic Acids Res.* **47,** D212–D220 (Jan. 2019).

11. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. en. *Nucleic Acids Res.* **43,** D261–9 (Jan. 2015).

12. King, A. N., Borkar, S. A., Samuels, D. J., Batz, Z., Bulock, L. L., Sadykov, M. R., Bayles, K. W. & Brinsmade, S. R. Guanine Limitation Results in CodY-Dependent and -Independent Alteration of Staphylococcus aureus Physiology and Gene Expression. *J. Bacteriol.* **200** (July 2018).

13. Pohl, K., Francois, P., Stenz, L., Schlink, F., Geiger, T., Herbert, S., Goerke, C., Schrenzel, J. & Wolz, C. CodY in Staphylococcus aureus: a regulatory link between metabolism and virulence gene expression. *J. Bacteriol.* **191,** 2953–2963 (May 2009).

14. Hove-Jensen, B., Andersen, K. R., Kilstrup, M., Martinussen, J., Switzer, R. L. & Willemoës, M. Phosphoribosyl Diphosphate (PRPP): Biosynthesis, Enzymology, Utilization, and Metabolic Significance. en. *Microbiol. Mol. Biol. Rev.* **81** (Mar. 2017).

15. Kofoed, E. M., Yan, D., Katakam, A. K., Reichelt, M., Lin, B., Kim, J., Park, S., Date, S. V., Monk, I. R., Xu, M., Austin, C. D., Maurer, T. & Tan, M.-W. De Novo Guanine Biosynthesis but Not the Riboswitch-Regulated Purine Salvage Pathway Is Required for Staphylococcus aureus Infection In Vivo. en. *J. Bacteriol.* **198,** 2001–2015 (July 2016).

16. Gaca, A. O., Colomer-Winter, C. & Lemos, J. A. *Many means to a common end: The intricacies of (p)ppGpp metabolism and its control of bacterial homeostasis* 2015.

17. Kriel, A., Bittner, A. N., Kim, S. H., Liu, K., Tehranchi, A. K., Zou, W. Y., Rendon, S., Chen, R., Tu, B. P. & Wang, J. D. Direct Regulation of GTP Homeostasis by (p)ppGpp: A Critical Component of Viability and Stress Resistance. *Mol. Cell* **48,** 231–241 (Oct. 2012).

18. Srivatsan, A. & Wang, J. D. Control of bacterial transcription, translation and replication by (p)ppGpp. *Curr. Opin. Microbiol.* **11,** 100–105 (2008).

19. Horsburgh, M. J., Wharton, S. J., Cox, A. G., Ingham, E., Peacock, S. & Foster, S. J. MntR modulates expression of the PerR regulon and superoxide resistance in Staphylococcus aureus through control of manganese uptake. en. *Mol. Microbiol.* **44,** 1269–1286 (June 2002).

20. Sadykov, M. R., Hartmann, T., Mattes, T. A., Hiatt, M., Jann, N. J., Zhu, Y., Ledala, N., Landmann, R., Herrmann, M., Rohde, H., Bischoff, M. & Somerville, G. A. CcpA coordinates central metabolism and biofilm formation in Staphylococcus epidermidis. en. *Microbiology* **157,** 3458–3468 (Dec. 2011).

21. Halsey, C. R., Lei, S., Wax, J. K., Lehman, M. K., Nuxoll, A. S., Steinke, L., Sadykov, M., Powers, R. & Fey, P. D. Amino acid catabolism in Staphylococcus aureus and the function of carbon catabolite repression. *MBio* **8** (2017).

22. Seidl, K., Müller, S., François, P., Kriebitzsch, C., Schrenzel, J., Engelmann, S., Bischoff, M. & Berger-Bächi, B. Effect of a glucose impulse on the CcpA regulon in Staphylococcus aureus. en. *BMC Microbiol.* **9,** 95 (May 2009).

23. Leiba, J., Hartmann, T., Cluzel, M.-E., Cohen-Gonsaud, M., Delolme, F., Bischoff, M. & Molle, V. A Novel Mode of Regulation of the *Staphylococcus aureus* Catabolite Control Protein A (CcpA) Mediated by Stk1 Protein Phosphorylation. *J. Biol. Chem.* **287,** 43607–43619 (Dec. 2012).

24. Bengert, P. & Dandekar, T. Riboswitch finder–a tool for identification of riboswitch RNAs. en. *Nucleic Acids Res.* **32,** W154–9 (July 2004).

25. Sause, W. E., Balasubramanian, D., Irnov, I., Copin, R., Sullivan, M. J., Sommerfield, A., Chan, R., Dhabaria, A., Askenazi, M., Ueberheide, B., Shopsin, B., van Bakel, H. & Torres, V. J. The purine biosynthesis regulator PurR moonlights as a virulence regulator in Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* (June 2019).

26. Cheung, G. Y. C., Wang, R., Khan, B. A., Sturdevant, D. E. & Otto, M. Role of the accessory gene regulator agr in community-associated methicillin-resistant Staphylococcus aureus pathogenesis. en. *Infect. Immun.* **79,** 1927–1935 (May 2011).

27. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. en. *Nat. Protoc.* **5,** 93–121 (Jan. 2010).

28. Seif, Y., Monk, J. M., Mih, N., Tsunemoto, H., Poudel, S., Zuniga, C., Broddrick, J., Zengler, K. & Palsson, B. O. A computational knowledge-base elucidates the response of Staphylococcus aureus to different media types. en. *PLoS Comput. Biol.* **15,** e1006644 (Jan. 2019).

29. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987 (May 2015).

30. Feist, A. M. & Palsson, B. O. The biomass objective function. en. *Curr. Opin. Microbiol.* **13,** 344–349 (June 2010).

31. Ding, Y., Liu, X., Chen, F., Di, H., Xu, B., Zhou, L., Deng, X., Wu, M., Yang, C.-G. & Lan, L. Metabolic sensor governing bacterial virulence in Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E4981–90 (Nov. 2014).

32. Horsburgh, M. J., Aish, J. L., White, I. J., Shaw, L., Lithgow, J. K. & Foster, S. J. sigmaB modulates virulence determinant expression and stress resistance: characterization of a functional rsbU strain derived from Staphylococcus aureus 8325-4. en. *J. Bacteriol.* **184,** 5457–5467 (Oct. 2002).

33. Basu, A., Shields, K. E., Eickhoff, C. S., Hoft, D. F. & Yap, M.-N. F. Thermal and nutritional regulation of ribosome hibernation in Staphylococcus aureus. en. *J. Bacteriol.* (Oct. 2018).

34. Morikawa, K., Takemura, A. J., Inose, Y., Tsai, M., Nguyen Thi, L. T., Ohta, T. & Msadek, T. Expression of a cryptic secondary sigma factor gene unveils natural competence for DNA transformation in Staphylococcus aureus. en. *PLoS Pathog.* **8,** e1003003 (Nov. 2012).

35. Miller, H. K., Carroll, R. K., Burda, W. N., Krute, C. N., Davenport, J. E. & Shaw, L. N. The extracytoplasmic function sigma factor σS protects against both intracellular and extracytoplasmic stresses in Staphylococcus aureus. en. *J. Bacteriol.* **194,** 4342–4354 (Aug. 2012).

36. Lorenz, U., Hüttinger, C., Schäfer, T., Ziebuhr, W., Thiede, A., Hacker, J., Engelmann, S., Hecker, M. & Ohlsen, K. The alternative sigma factor sigma B of Staphylococcus aureus modulates virulence in experimental central venous catheter-related infections. en. *Microbes Infect.* **10,** 217–223 (Mar. 2008).

37. Tuchscherr, L., Bischoff, M., Lattar, S. M., Noto Llana, M., Pförtner, H., Niemann, S., Geraci, J., Van de Vyver, H., Fraunholz, M. J., Cheung, A. L., Herrmann, M., Völker, U., Sordelli, D. O., Peters, G. & Löffler, B. Sigma Factor SigB Is Crucial to Mediate Staphylococcus aureus Adaptation during Chronic Infections. en. *PLoS Pathog.* **11,** e1004870 (Apr. 2015).

38. Senn, M. M., Giachino, P., Homerova, D., Steinhuber, A., Strassner, J., Kormanec, J., Flückiger, U., Berger-Bächi, B. & Bischoff, M. Molecular analysis and organization of the sigmaB operon in Staphylococcus aureus. en. *J. Bacteriol.* **187,** 8006–8019 (Dec. 2005).

39. Tamber, S., Schwartzman, J. & Cheung, A. L. Role of PknB kinase in antibiotic resistance and virulence in community-acquired methicillin-resistant Staphylococcus aureus strain USA300. *Infect. Immun.* **78,** 3637–3646 (Aug. 2010).

40. Mäder, U., Nicolas, P., Depke, M., Pané-Farré, J., Debarbouille, M., van der Kooi-Pol, M. M., Guérin, C., Dérozier, S., Hiron, A., Jarmer, H., Leduc, A., Michalik, S., Reilman, E., Schaffer, M., Schmidt, F., Bessières, P., Noirot, P., Hecker, M., Msadek, T., Völker, U. & van Dijl, J. M. Staphylococcus aureus Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions. en. *PLoS Genet.* **12,** e1005962 (Apr. 2016).

41. Burda, W. N., Miller, H. K., Krute, C. N., Leighton, S. L., Carroll, R. K. & Shaw, L. N. Investigating the genetic regulation of the ECF sigma factor σS in Staphylococcus aureus. en. *BMC Microbiol.* **14,** 280 (Nov. 2014).

42. Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. *Cell systems* **2,** 260–271 (Apr. 2016).

43. Jenul, C. & Horswill, A. R. Regulation of Staphylococcus aureus Virulence. en. *Microbiol Spectr* **6** (Feb. 2018).

44. Liu, Q., Yeo, W.-S. & Bae, T. The SaeRS Two-Component System of Staphylococcus aureus. en. *Genes* **7** (Oct. 2016).

45. Goncheva, M. I., Flannagan, R. S., Sterling, B. E., Laakso, H. A., Friedrich, N. C., Kaiser, J. C., Watson, D. W., Wilson, C. H., Sheldon, J. R., McGavin, M. J., Kiser, P. K. & Heinrichs, D. E. Stress-induced inactivation of the Staphylococcus aureus purine biosynthesis repressor leads to hypervirulence. en. *Nat. Commun.* **10,** 775 (Feb. 2019).

46. Novick, R. P., Projan, S. J., Kornblum, J., Ross, H. F., Ji, G., Kreiswirth, B., Vandenesch, F. & Moghazeh, S. The agr P2 operon: an autocatalytic sensory transduction system in Staphylococcus aureus. en. *Mol. Gen. Genet.* **248,** 446–458 (Aug. 1995).

47. Novick, R. P. & Geisinger, E. Quorum sensing in staphylococci. en. *Annu. Rev. Genet.* **42,** 541–564 (2008).

48. Regassa, L. B. & Betley, M. J. Alkaline pH decreases expression of the accessory gene regulator (agr) in Staphylococcus aureus. en. *J. Bacteriol.* **174,** 5095–5100 (Aug. 1992).

49. Weinrick, B., Dunman, P. M., McAleese, F., Murphy, E., Projan, S. J., Fang, Y. & Novick, R. P. Effect of mild acid on gene expression in Staphylococcus aureus. en. *J. Bacteriol.* **186,** 8407–8423 (Dec. 2004).

50. Regassa, L. B., Novick, R. P. & Betley, M. J. Glucose and nonmaintained pH decrease expression of the accessory gene regulator (agr) in Staphylococcus aureus. en. *Infect. Immun.* **60,** 3381–3388 (Aug. 1992).

51. Joo, H.-S., Chan, J. L., Cheung, G. Y. C. & Otto, M. Subinhibitory concentrations of protein synthesis-inhibiting antibiotics promote increased expression of the agr virulence regulator and production of phenol-soluble modulin cytolysins in community-associated methicillin-resistant Staphylococcus aureus. en. *Antimicrob. Agents Chemother.* **54,** 4942–4944 (Nov. 2010).

52. Pöhlmann-Dietze, P., Ulrich, M., Kiser, K. B., Döring, G., Lee, J. C., Fournier, J. M., Botzenhart, K. & Wolz, C. Adherence of Staphylococcus aureus to endothelial cells: influence of capsular polysaccharide, global regulator agr, and bacterial growth phase. en. *Infect. Immun.* **68,** 4865–4871 (Sept. 2000).

53. Lee, J. C., Takeda, S., Livolsi, P. J. & Paoletti, L. C. Effects of in vitro and in vivo growth conditions on expression of type 8 capsular polysaccharide by Staphylococcus aureus. en. *Infect. Immun.* **61,** 1853–1858 (May 1993).

54. Verkaik, N. J., Benard, M., Boelens, H. A., de Vogel, C. P., Nouwen, J. L., Verbrugh, H. A., Melles, D. C., van Belkum, A. & van Wamel, W. J. B. Immune evasion cluster-positive bacteriophages are highly prevalent among human Staphylococcus aureus strains, but they are not essential in the first stages of nasal colonization. en. *Clin. Microbiol. Infect.* **17,** 343–348 (Mar. 2011).

55. Bae, T., Baba, T., Hiramatsu, K. & Schneewind, O. Prophages of Staphylococcus aureus Newman and their contribution to virulence. en. *Mol. Microbiol.* **62,** 1035–1047 (Nov. 2006).

56. Jones, M. B., Montgomery, C. P., Boyle-Vavra, S., Shatzkes, K., Maybank, R., Frank, B. C., Peterson, S. N. & Daum, R. S. Genomic and transcriptomic differences in community ac-

quired methicillin resistant Staphylococcus aureus USA300 and USA400 strains. en. *BMC Genomics* **15,** 1145 (Dec. 2014).

57. Malachowa, N., Kobayashi, S. D., Sturdevant, D. E., Scott, D. P. & DeLeo, F. R. Insights into the Staphylococcus aureus-host interface: global changes in host and pathogen gene expression in a rabbit skin infection model. en. *PLoS One* **10,** e0117713 (Feb. 2015).

58. Malachowa, N., Whitney, A. R., Kobayashi, S. D., Sturdevant, D. E., Kennedy, A. D., Braughton, K. R., Shabb, D. W., Diep, B. A., Chambers, H. F., Otto, M. & DeLeo, F. R. Global changes in Staphylococcus aureus gene expression in human blood. en. *PLoS One* **6,** e18617 (Apr. 2011).

59. Peterson, M. M., Mack, J. L., Hall, P. R., Alsup, A. A., Alexander, S. M., Sully, E. K., Sawires, Y. S., Cheung, A. L., Otto, M. & Gresham, H. D. Apolipoprotein B Is an innate barrier against invasive Staphylococcus aureus infection. en. *Cell Host Microbe* **4,** 555–566 (Dec. 2008).

60. Hall, P. R., Elmore, B. O., Spang, C. H., Alexander, S. M., Manifold-Wheeler, B. C., Castleman, M. J., Daly, S. M., Peterson, M. M., Sully, E. K., Femling, J. K., Otto, M., Horswill, A. R., Timmins, G. S. & Gresham, H. D. Nox2 modification of LDL is essential for optimal apolipoprotein B-mediated control of agr type III Staphylococcus aureus quorum-sensing. en. *PLoS Pathog.* **9,** e1003166 (Feb. 2013).

61. Tonyushkina, K. & Nichols, J. H. Glucose meters: a review of technical challenges to obtaining accurate results. en. *J. Diabetes Sci. Technol.* **3,** 971–980 (July 2009).

62. Connolly, J., Boldock, E., Prince, L. R., Renshaw, S. A., Whyte, M. K. & Foster, S. J. Identification of Staphylococcus aureus Factors Required for Pathogenicity and Growth in Human Blood. en. *Infect. Immun.* **85** (Nov. 2017).

63. O'Riordan, K. & Lee, J. C. Staphylococcus aureus capsular polysaccharides. *Clin. Microbiol. Rev.* **17,** 218–234 (Jan. 2004).

64. Stubbendieck, R. M., May, D. S., Chevrette, M. G., Temkin, M. I., Wendt-Pienkowski, E., Cagnazzo, J., Carlson, C. M., Gern, J. E. & Currie, C. R. Competition among Nasal Bacteria Suggests a Role for Siderophore-Mediated Interactions in Shaping the Human Nasal Microbiota. en. *Appl. Environ. Microbiol.* **85** (May 2019).

65. Minch, K. J., Rustad, T. R., Peterson, E. J. R., Winkler, J., Reiss, D. J., Ma, S., Hickey, M., Brabant, W., Morrison, B., Turkarslan, S., Mawhinney, C., Galagan, J. E., Price, N. D., Baliga, N. S. & Sherman, D. R. The DNA-binding network of Mycobacterium tuberculosis. en. *Nat. Commun.* **6,** 5829 (Jan. 2015).

66. Choe, D., Szubin, R., Dahesh, S., Cho, S., Nizet, V., Palsson, B. & Cho, B.-K. Genome-scale analysis of Methicillin-resistant Staphylococcus aureus USA300 reveals a tradeoff between pathogenesis and drug resistance. en. *Sci. Rep.* **8,** 2215 (Feb. 2018).

67. Poudel, S., Tsunemoto, H., Meehan, M., Szubin, R., Olson, C. A., Lamsa, A., Seif, Y., Dillon, N., Vrbanac, A., Sugie, J., Dahesh, S., Monk, J. M., Dorrestein, P. C., Pogliano,

J., Knight, R., Nizet, V., Palsson, B. O. & Feist, A. M. Characterization of CA-MRSA TCH1516 exposed to nafcillin in bacteriological and physiological media. en. *Sci Data* **6,** 43 (Apr. 2019).

68. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. en. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 2896–2901 (Mar. 1999).

69. Highlander, S. K., Hulten, K. G., Qin, X., Jiang, H., Yerrapragada, S., Mason, E. O., Shang, Y., Williams, T. M., Fortunov, R. M., Liu, Y., Igboeli, O., Petrosino, J., Tirumalai, M., Uzman, A., Fox, G. E., Cardenas, A. M., Muzny, D. M., Hemphill, L., Ding, Y., Dugan, S., Blyth, P. R., Buhay, C. J., Dinh, H. H., Hawes, A. C., Holder, M., Kovar, C. L., Lee, S. L., Liu, W., Nazareth, L. V., Wang, Q., Zhou, J., Kaplan, S. L. & Weinstock, G. M. Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. *BMC Microbiol.* **7,** 99 (Nov. 2007).

70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (Mar. 2012).

71. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. en. *Bioinformatics* **31,** 166–169 (Jan. 2015).

72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (Dec. 2014).

73. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

74. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. en. *Neural Netw.* **13,** 411–430 (May 2000).

75. D'Agostino, R. B. & Belanger, A. A Suggestion for Using Powerful and Informative Tests of Normality. *Am. Stat.* **44,** 316–321 (1990).

76. Ravcheev, D. A., Best, A. A., Tintle, N., Dejongh, M., Osterman, A. L., Novichkov, P. S. & Rodionov, D. A. Inference of the transcriptional regulatory network in Staphylococcus aureus by integration of experimental and genomics-based evidence. *J. Bacteriol.* **193,** 3228–3240 (July 2011).

77. Boyle-Vavra, S., Yin, S., Jo, D. S., Montgomery, C. P. & Daum, R. S. VraT/YvqF is required for methicillin resistance and activation of the VraSR regulon in Staphylococcus aureus. en. *Antimicrob. Agents Chemother.* **57,** 83–95 (Jan. 2013).

78. Kuroda, M., Kuroda, H., Oshima, T., Takeuchi, F., Mori, H. & Hiramatsu, K. Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in Staphylococcus aureus. en. *Mol. Microbiol.* **49,** 807–821 (Aug. 2003).

79. Delauné, A., Dubrac, S., Blanchet, C., Poupel, O., Mäder, U., Hiron, A., Leduc, A., Fitting, C., Nicolas, P., Cavaillon, J.-M., Adib-Conquy, M. & Msadek, T. The WalKR system controls major staphylococcal virulence genes and is involved in triggering the host inflammatory response. *Infect. Immun.* **80,** 3438–3453 (Oct. 2012).

80. Falord, M., Mäder, U., Hiron, A., Débarbouillé, M. & Msadek, T. Investigation of the Staphylococcus aureus GraSR regulon reveals novel links to virulence, stress response and cell wall signal transduction pathways. en. *PLoS One* **6,** e21323 (July 2011).

81. Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing Group, Lau, B. Y., Hoad, G., Winsor, G. L. & Brinkman, F. S. L. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. en. *Nucleic Acids Res.* **45,** W30–W35 (July 2017).

82. Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D. S. PHASTER: a better, faster version of the PHAST phage search tool. en. *Nucleic Acids Res.* **44,** W16–21 (July 2016).

83. Ermolaeva, M. D., White, O. & Salzberg, S. L. Prediction of operons in microbial genomes. en. *Nucleic Acids Res.* **29,** 1216–1221 (Mar. 2001).

84. Pertea, M., Ayanbule, K., Smedinghoff, M. & Salzberg, S. L. OperonDB: a comprehensive database of predicted operons in microbial genomes. en. *Nucleic Acids Res.* **37,** D479–82 (Jan. 2009).

85. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. en. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).

86. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. en. *Nucleic Acids Res.* **37,** W202–8 (July 2009).

87. Kiliç, S., White, E. R., Sagitova, D. M., Cornish, J. P. & Erill, I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. en. *Nucleic Acids Res.* **42,** D156–60 (Jan. 2014).

88. Eckweiler, D., Dudek, C.-A., Hartlich, J., Brötje, D. & Jahn, D. PRODORIC2: the bacterial gene regulation database in 2018. en. *Nucleic Acids Res.* **46,** D320–D326 (Jan. 2018).

89. Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S. & Dubchak, I. RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. en. *Nucleic Acids Res.* **35,** D407–12 (Jan. 2007).

90. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. en. *Genome Biol.* **8,** R24 (2007).

91. Hammer, N. D., Schurig-Briccio, L. A., Gerdes, S. Y., Gennis, R. B. & Skaar, E. P. CtaM Is Required for Menaquinol Oxidase aa3 Function in Staphylococcus aureus. en. *MBio* **7** (July 2016).

92. Nuxoll, A. S., Halouska, S. M., Sadykov, M. R., Hanke, M. L., Bayles, K. W., Kielian, T., Powers, R. & Fey, P. D. CcpA regulates arginine biosynthesis in Staphylococcus aureus through repression of proline catabolism. en. *PLoS Pathog.* **8,** e1003033 (Nov. 2012).

93. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. en. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

94. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. en. *PLoS Comput. Biol.* **4,** e1000082 (May 2008).

95. Hyduke, D. R., Lewis, N. E. & Palsson, B. Ø. Analysis of omics data with genome-scale models of metabolism. en. *Mol. Biosyst.* **9,** 167–174 (Feb. 2013).

96. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. en. *Bioinformatics* **20,** 307–315 (Feb. 2004).

Chapter 3

Coupling of CcpA and CodY activities coordinates carbon and nitrogen metabolism associated gene expression in *S. aureus* USA300 strains

3.1 Abstract

The complex crosstalk between metabolism and gene regulatory networks makes it difficult to untangle individual constituents and study their precise roles and interactions. To address this issue, we modularized the transcriptional regulatory network (TRN) of the *Staphylococcus* *aureus* USA300 strain by applying Independent Component Analysis (ICA) to 385 RNA sequencing samples. We then combined the modular TRN model with a metabolic model to study the regulation of carbon and amino acid metabolism. Our analysis showed that regulation of central carbon metabolism by CcpA and central nitrogen metabolism by CodY are closely coordinated. *S. aureus*, in general, increases the expression of CodY-regulated genes in the presence of preferred carbons such as glucose. This transcriptional coordination was corroborated by simulations with metabolic models that also showed increased amino acid biosynthesis in the presence of glucose. Further, CodY and CcpA cooperatively regulate the expression of ribosome hibernation promoting factor, thus linking metabolic cues with translation. In line with this hypothesis, expression of CodY regulated genes is tightly correlated with expression of genes encoding ribosomal proteins. Together, we propose a coarse-grained model where expression of *S. aureus* genes encoding enzymes that control carbon flux and nitrogen flux through the system is coregulated with expression of translation machinery to modularly control protein synthesis. While this work focuses on three key regulators, the full TRN model we present contains 76 total independently modulated sets of genes, each with the potential to uncover other complex regulatory structures and interactions.

3.2 Introduction

Metabolism plays an integral role in infection and antimicrobial resistance (AMR) in the leading human bacterial pathogen Staphylococcus aureus. Metabolic requirements specific to infection, intracellular persistence, biofilm formation, and colonization are rapidly being uncovered[1–6]. Furthermore, the central role of metabolism in AMR and persistence is also coming into view, adding to the complexity of known AMR mechanisms[7–9]. The complex metabolic circuits and responses underlying these phenomena are nevertheless difficult to unravel. Even relatively well understood systems such as S. aureus central carbon metabolism can be difficult to fully map as they are layered with multiple levels of gene regulation, post-translational and biochemical controls, and unexpected molecular interactions[1, 10–12]. Some of these complexities can be captured by genome-scale metabolic models (GEMs) that allow rapid query of metabolic complexities through simulations of metabolic flux states, knock-out experiments, multi-strain metabolic comparisons, and more[13, 14]. Alternatively, coarse-grained modeling of metabolism attempts to peer beyond the detailed complexity and discover the general principles governing the system. In the present work, we took guidance from the coarse-grained model proposed in *Escherichia coli* coupled with, genome scale analyses of S. aureus transcriptional regulation and metabolism to uncover similar staphylococcal system that balances resource allocation between carbon and nitrogen metabolism[15–17].

Biological trade-offs represent an optimization frontier, where the cell must strike a balance between its multiple objectives and their limitations[15, 18]. Signatures of these balancing acts can be found in transcriptomes and become apparent when their architecture is viewed at systems level[19]. We previously described one such trade-off and its transcriptional imprint using independent data sets from Gram-negative *E. coli* and Gram-positive *S. aureus*- in which a balance was struck between genes regulated by stress associated sigma factors and growth associated translation machinery[20, 21]. This trade-off was observed in independent data sets in both gram-negative *E. coli* and gram-positive *S. aureus.* Here, we expand significantly beyond those observations to describe a trade-off between carbon and nitrogen metabolism in strains of the globally disseminated, hypervirulent *S. aureus* USA300 lineage.

We first greatly expanded on our previously published transcriptional regulatory model of USA300 strains to incorporate all publicly available RNA sequencing data from the Sequence Reads Archive (SRA)[21]. Models were then generated using independent component analysis (ICA) that calculates independently modulated sets of genes (iModulons) and their activities present in the input RNA sequencing samples. iModulons represent sources of signals in the expression data, with transcriptional regulators being the most common source. Our model showed that the activities of two global metabolic regulators, CcpA and CodY, which play critical roles in central carbon and nitrogen metabolism respectively, are negatively correlated to one another. This negative correlation pointed to condition-specific reallocation of resources towards different metabolic subsystems. GEMs fitted with metabolomics data confirmed the inferences made from the transcriptomic data. Furthermore, GEMs revealed specific metabolic intersections including glutamate dehydrogenase and the folate cycle where coordination of metabolism by the two regulators is required for optimal biomass production. Placing genes from CodY and CcpA- associated iModulons onto the metabolic map demonstrated that they did not share any metabolic reactions, but coregulated the expression of a gene encoding ribosome hibernation factor. In light of these observations, we propose a model whereby CcpA and CodY coordinate gene expression for carbon metabolism, nitrogen metabolism and translation, thus coordinating protein production at specific stages.

3.3 Results

Expanding the USA300 iModulons using RNA-sequencing data from SRA database

Our previous work outlined 29 iModulons for USA300 strains that were generated from 108 in-house RNA-sequencing data[21]. To expand that model, we queried Sequence Reads Archive (SRA) for all available USA300 specific RNA-sequencing data (Figure B.1) and combined it with 64 newly generated samples. Of the 576 sequencing samples available, 385 passed the stringent QC/QA pipeline and were therefore incorporated into the new model (see Methods). The final set of samples contained data from multiple at least 7 different USA300 isolates, 4 growth phases (exponential, stationary, biofilm and infection) and 10 base medium (Figure B.2).

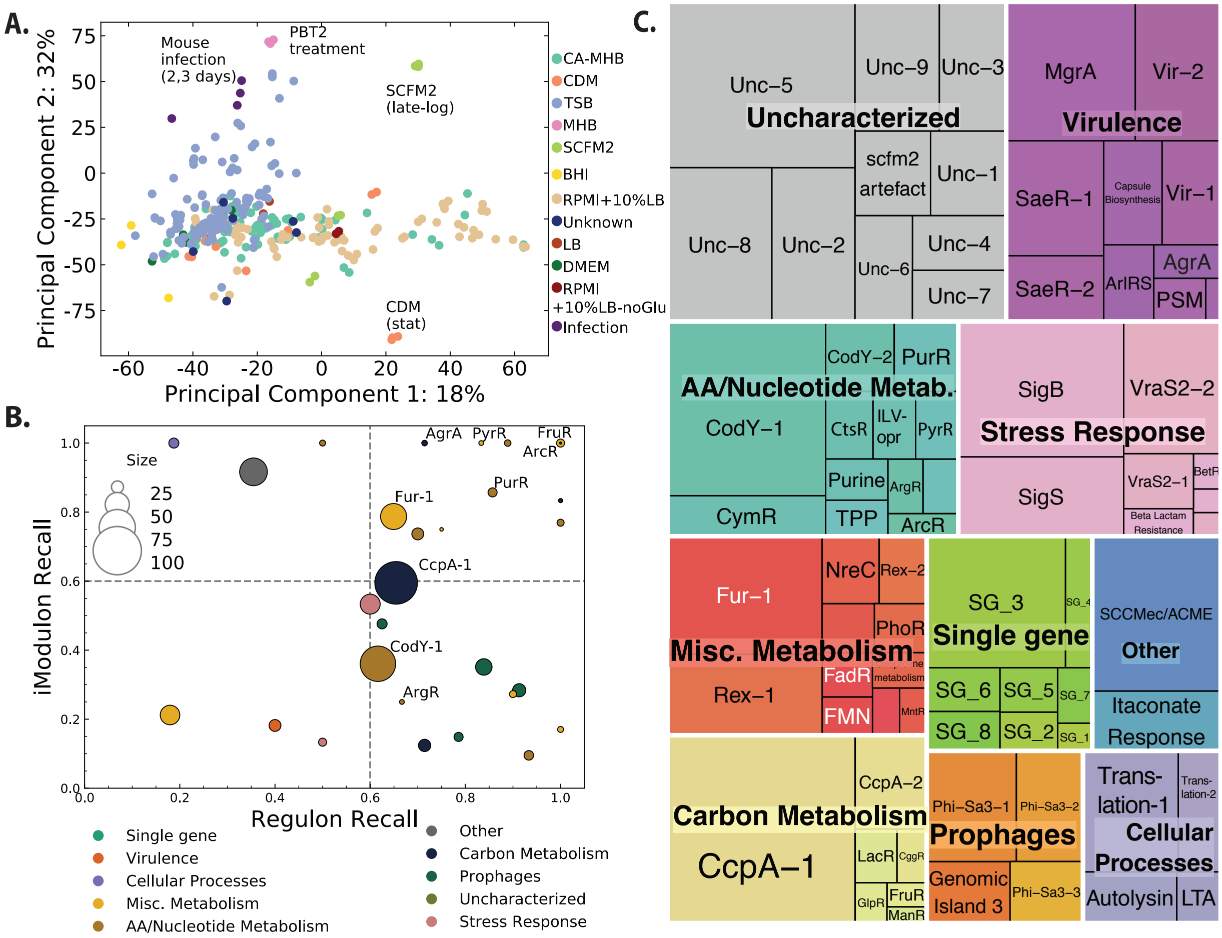


Figure 3.1: The updated iModulons for USA300 strains. (a) 385 RNA-sequencing samples from diverse growth conditions were used to generate the expanded USA300 iModulons. The samples were normalized to project specific control conditions to reduce signal from batch effect. (b) iModulons were labeled based on significant association with other published regulons. (c) Full iModulon names, size (gene content) and types in the current model after manual curation.

Before applying ICA, we normalized the log transformed Transcripts per Million (log-TPM) data to a project specific control condition. This reduced batch specific variation in the data and reduced iModulons not associated with biological signals. Principal component analysis of the log-TPM data showed that normalized samples tended to cluster with media types and growth phases rather than by project (Figure 3.1a). For example, data from *S. aureus* grown to late-log phase in SCFM2 (Synthetic Cystic Fibrosis Sputum Medium 2) and to stationary phase in CDM did not cluster together, despite being from the same project.

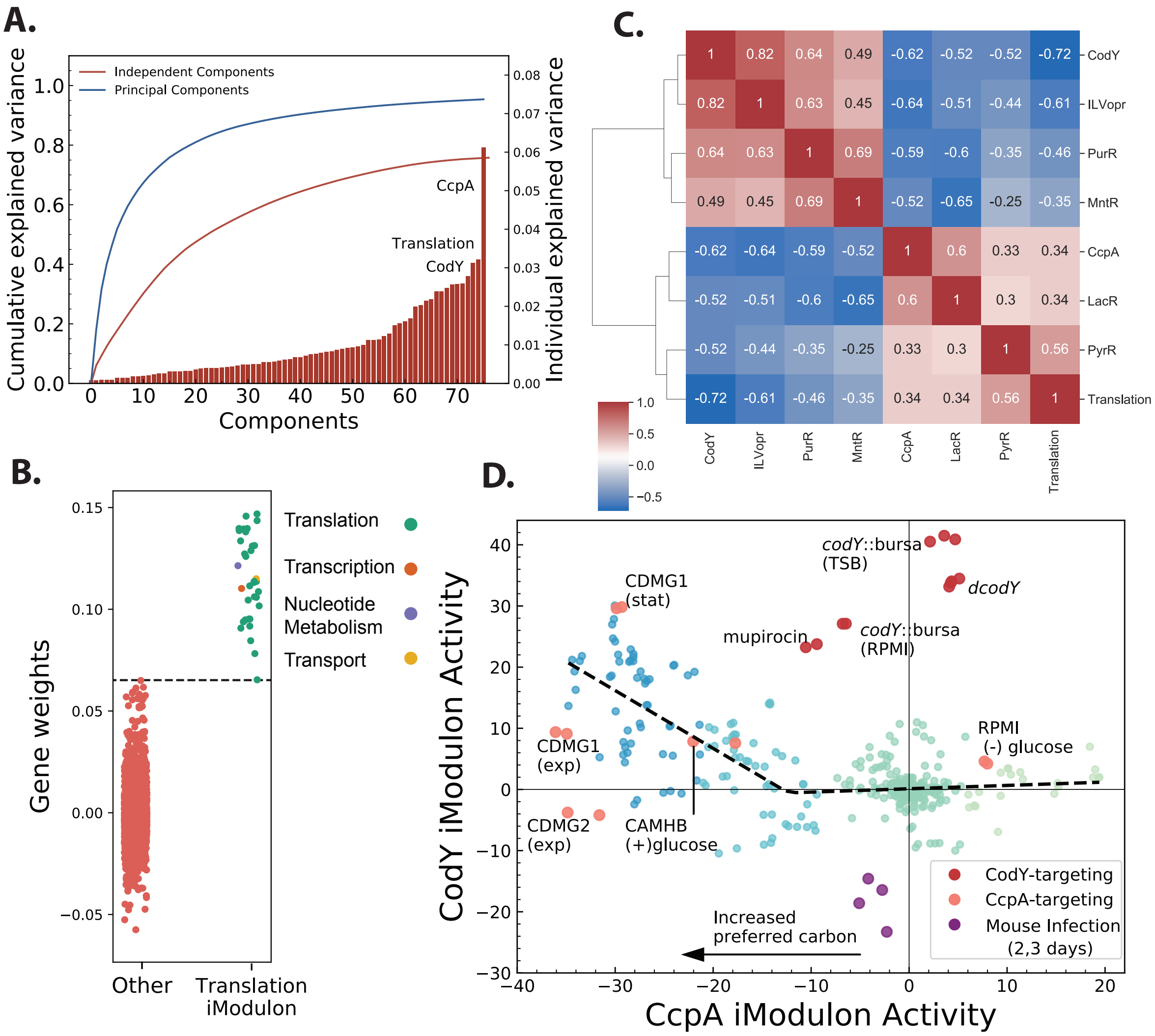
Application of ICA to this normalized log-TPM data extracted 76 independent components and genes with high absolute weightings within each components were assigned to a corresponding iModulon. These enriched iModulon genes were then compared with existing literature of predicted regulons in *S. aureus*. Those iModulons that had significant overlap with other predicted regulons were named after the associated regulator (Figure 3.1b). Lastly, some iModulons with no known regulators, but associated other biological processes (e.g. prophages, translation) were manually curated. In all, we were able to label 60 of the 76 iModulons with either a regulator or a biological process (Figure 3.1c). In addition to the structure of the iModulon, the activities of each of the 76 iModulons in the 385 input samples were also calculated. The activity represents the role each iModulon (and the associated regulator if known) in shaping the role of transcriptome in the given sample. Higher iModulon activity represents higher expression level of genes with positive weightings in the iModulon and lower expression of genes with negatively weighted genes.

CcpA and CodY iModulon activities highlight balance of carbon and nitrogen metabolism

Cumulatively, the 70 iModulons captured 70% of the variance in the input transcriptomic data. The CodY-2, CcpA-3 (henceforth referred to as simply CodY and CcpA) and Translation iModulons had the highest explained variance (Figure 3.2a). CcpA is the catabolite repressor protein in firmicutes that represses genes involved in alternate carbon utilization as well as other central carbon metabolic pathways such as the Tricarboxylic acid (TCA) cycle in the presence of high concentrations of glucose[22]. CodY, on the other hand, globally represses the genes required for amino acid biosynthesis in response to high branched chain amino acid (BCAA) or GTP concentrations. Lastly, the Translation iModulon almost entirely consists of ribosomal genes (e.g. rplK, rplA etc.) and genes involved in translation such as infA and fusA which encode translation initiation factor IF-1 and elongation factor G respectively. This iModulon has been enriched in almost all bacteria and archaea for which iModulons have been calculated [20, 23–26].

Interestingly, activities of these three iModulons were highly correlated across all samples (Figure 3.2b). Along with CodY, CcpA, and Translation iModulons, activities of ILVopr(iModulon containing the operon with isoleucine, leucine, and valine biosynthesis genes), MntR, LacR PyrR and PurR iModulons were also highly correlated (Figure B.3). Correlation of CcpA with LacR simply reflects the catabolite repression of lactose utilization genes by the regulator CcpA. Similarly, ILV operon is regulated globally by CodY and locally by leucine attenuator [27]. This multi-layer regulation likely explains why this operon formed its own iModulon whose activity was closely correlated with CodY. MntR iModulon contains genes required for manganese uptake and its coordinated activity with CcpA confirms the association of manganese concentration with glycolytic flux[28].

The correlated activity of CcpA and CodY iModulons suggested that *S. aureus* carefully coordinates its central carbon and nitrogen metabolism (Figure 3.2c). Close examination of the activities of these two iModulons showed a biphasic relationship. In conditions with preferred carbon sources and therefore low CcpA iModulon activity, CodY activity generally increased. This



	CodY	ILVopr	PurR	MntR	CcpA	LacR	PyrR	Translation
CodY	1	0.82	0.64	0.49	-0.62	-0.52	-0.52	-0.72
ILVopr	0.82	1	0.63	0.45	-0.64	-0.51	-0.44	-0.61
PurR	0.64	0.63	1	0.69	-0.59	-0.6	-0.35	-0.46
MntR	0.49	0.45	0.69	1	-0.52	-0.65	-0.25	-0.35
CcpA	-0.62	-0.64	-0.59	-0.52	1	0.6	0.33	0.34
LacR	-0.52	-0.51	-0.6	-0.65	0.6	1	0.3	0.34
PyrR	-0.52	-0.44	-0.35	-0.25	0.33	0.3	1	0.56
Translation	-0.72	-0.61	-0.46	-0.35	0.34	0.34	0.56	1

Figure 3.2:

(a) Explained variance of each of the iModulons; CcpA, Translation and CodY iModulons explain the most variance in the transcriptome data. (b) Correlation between various metabolic iModulons highlights coordination of gene expression between various metabolic subsystems. (c) Activity of CcpA and CodY iModulons across all USA300 samples. Inactivation of CodY does not alter CcpA activity but decrease in CcpA activity leads to increase in CodY activity. This asymmetric relationship suggests that CcpA works upstream of CodY.

effect was observed when glucose was added to both complex- Cation-Adjusted Mueller Hinton Broth (CA-MHB)- and to a defined- Chemically Defined Medium(CDM1)- medium. Other conditions without explicitly controlled glucose levels that showed low CcpA activity still had concomitant high CodY activity, suggesting that this effect was not glucose specific. In conditions with already low CodY activity however, removal of glucose (RPMI (-) glucose; substituted with maltose) did not lead to further change in CodY activity, creating the second phase of the trade-off plane.

On the other hand, increase in CodY iModulon activity did not necessarily lead to decrease in CcpA activity (Figure 3.2c; red markers). Samples from codY interrupted strains in several different projects showed minimal effect on CcpA iModulon activity. These samples fell well outside of the CcpA-CodY trade-off line (Figure 3.2c; grey dashed lines). Similar effects can also be observed in samples treated with sub-inhibitory concentration of mupirocin. Mupirocin activates the stringent response in *S. aureus* which leads to conversion of GTP to ppgpp and subsequent derepression of CodY regulon[29]. As change in CcpA activity leads to change in CodY activity but not necessarily vice-versa, this data suggests that CcpA works 'upstream' of CodY.

3.3.1 Metabolic modeling confirms CcpA and CodY iModulon association

We used a previously published USA300 strain specific genome scale metabolic model (GEM) to independently confirm the metabolic interaction between CodY and CcpA[30]. GEMs are curated and mathematically formulated models of an organism's metabolism that can be used to simulate, study and design the metabolic pathways using a wide range of Constraints Based Analysis and Reconstruction (COBRA) tools[14, 31].

One such method, parsimonious Flux Balance Analysis (pFBA), can be used to calculate metabolic flux state that optimizes a phenotype while minimizing total metabolic flux in a given condition[13, 32]. Here, we used pFBA to determine the metabolic flux states that maximize *S. aureus* biomass production given the measured uptake and secretion rates of various amino acids and sugars in Chemically Defined Medium (CDM) and CDM + glucose (CDMG)[10]. In agreement with increased CodY iModulon activity in CDMG, total flux through reactions catalyzed by enzymes that are encoded in CodY iModulon genes ("CodY reactions" for short), doubled from 3 mmol/gDW/hr to 6 mmol/gDW/hr in presence of glucose (Figure 3.3a). A small decrease in CcpA reactions was also observed.

pFBA however, gives an exact optimal solution and therefore does not account for variations or errors in input uptake data. We addressed this issue by sampling the CDM and CDMG specific models which give distribution of feasible fluxes in each of the respective conditions. We then mapped the flux distribution to various amino acid biosynthetic pathways. For simple interpretation, we excluded amino acids that serve as intermediates for biosynthesis of other amino acids (e.g. glutamine, glutamate and serine) and included only those amino acid for which unique biosynthetic pathways could be defined (see Materials and Methods). Confirming pFBA analysis, 5 out of the 6 amino acid biosynthetic pathways had increased flux in CDMG when compared to CDM (Figure 3.3b). The results of these two TRN agnostic metabolic modeling methods are in agreement with our observation that CodY iModulon activity increases in presence of glucose.

CcpA and CodY reactions are coordinated at metabolic intersections

CcpA and CodY contained 110 and 86 genes respectively, with most genes involved in central carbon and amino acid metabolism. Despite the large iModulon sizes and close metabolic proximity of the regulated genes, the two iModulons did not share any genes encoding metabolic enzymes. The correlation in activity however, suggested that CcpA reaction and CodY reactions must coordinate at metabolic level. Using USA300 GEM, we looked for this coordination at the metabolite intersection of CcpA and CodY reactions i.e. metabolites that are involved in both CcpA and CodY reactions.

We found these intersection metabolites by systematically looking for all metabolites in USA300 GEM that can be found in both CodY and CcpA reactions. After taking out 'non-specific' metabolites and cofactors (e.g. ATP, H2O, NADH etc), we were left with 22 intersection metabolites (Table B.1). While some of these intersections like pyruvate, glutamate and oxaloacetate were expected as they play a crucial role in both carbon and nitrogen metabolism, other intersection metabolites like sl2a6 and tetrahydrofolate (THF) are less understood in the context of this trade-off. To further understand how change in simulated flux through CcpA and CodY reactions in CDM and CDMG altered these key metabolic intersections we mapped the pFBA solution fluxes from each media to the reactions around two of these intersections - glutamate and methylTHF.

The glutamate-alpha ketoglutarate (αkg) link is a closely studied intersection in *S. aureus* that connects amino acid and central carbon metabolism [5, 10]. The main enzyme at the intersection, glutamate dehydrogenase (GLUDy) reversibly iterconverts αkg and glutamate and is encoded by gudB gene, a constituent of the CcpA iModulon. However, this interconversion also acts as an amine group donor or acceptor to 3 CcpA reactions and 8 CodY reactions (Table B.1). In glucose free CDM, pFBA solution agreed with previous observation showing proline is converted to αkg via glutamate and eventually fuels gluconeogenesis[10] (Figure B.3). However, in CDMG, the flux through GLUDy changes direction and catalyzes conversion of αkg to glutamate

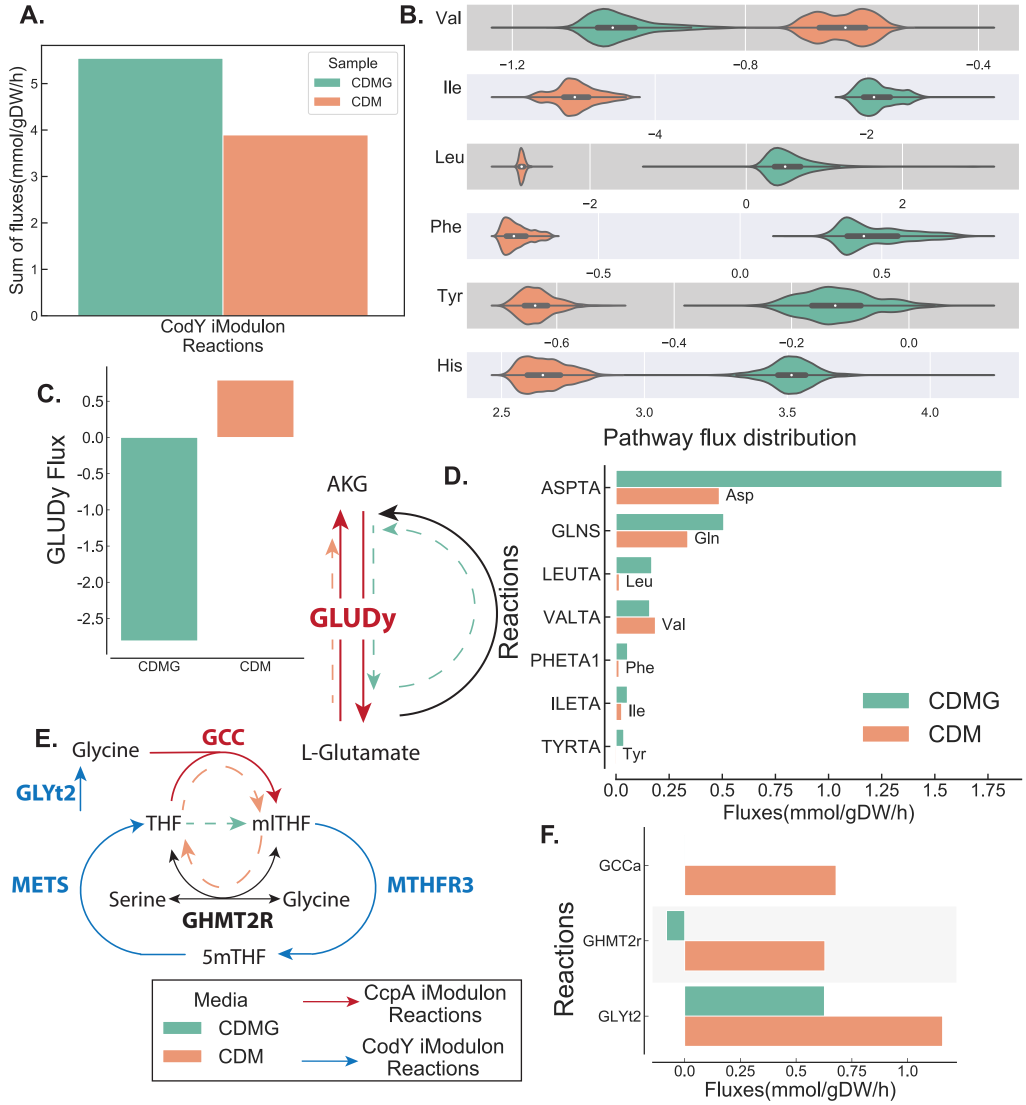


Figure 3.3: (a) Sum of sampled fluxes through CodY and CcpA reaction shows increased flux through CodY in CDMG. (b) Sampled fluxes through several amino acid biosynthesis pathways also show increased flux in CDMG. (c) Flux through GLUDy reaction changes direction when glucose is added. (d) Amino acids generated by accepting amine groups from L-glutamate. L-glutamate is converted to akg in the process and regenerated by GLUDy. (e) Metabolic map of folate cycle where CcpA and CodY regulated metabolism intersect. (f) Flux through reactions in folate cycle in CDM and CDMG.

instead(Figure 3.3c). This makes up 98% of total flux that consumes αkg. The glutamate in turn acts as an amine group donor for biosynthesis of various amino acids and accounts for 80% of total flux generating αkg in CDMG (Figure 3.3d). pFBA solution of this intersection therefore shows that in absence of glucose, GLUDy reaction converts glutamate to αkg to fuel gluconeogenesis but in the presence of glucose it converts αkg to glutamate to fuel amino acid biosynthesis.

The folate cycle represents another metabolic intersection of CcpA and CodY reactions. The folate cycle is required for one carbon metabolism, nucleotide biosynthesis and amino acid metabolism and the pathway leading up to the cycle is the target of sulfonamide class antibiotics[33]. The cycle consisted of 2 CodY reactions- MTHFR3 and METS (methionine synthase) - and one CcpA reaction- GCCabc (glycine cleavage complex) (Figure 3.3e). In CDM, tetrahydrofolate (THF) is converted to 5,10-methylenetetrahydrofolate (mlTHF) by GCCabc reaction which cleaves glycine in the process (Figure 3.3f). THF is then regenerated from mlTHF by GHMT2r reaction which also consumes glycine and generates serine. This consumption of glycine in folate cycle by CcpA reaction is coupled with increased transport of glycine by CodY regulated GLYt2. However, in CDMG where CcpA iModulon activity is low, there is no flux through the CcpA reaction, GCCabc. Instead, GHMT2r runs in 'reverse' to convert THF from mlTHF consuming serine and generating glycine instead. Together, combining iModulon structure with metabolic simulation demonstrates how despite not sharing any genes at regulatory level,*S. aureus* coordinates flux through CcpA and CodY iModulon reactions at these key metabolic intersections.

3.3.2 CcpA and CodY iModulons are coordinated with Translation iModulon

While CcpA and CodY iModulons do not share any metabolic genes, *hpf*, which encodes ribosomal hibernation promoting factors (HPF), was enriched in both iModulons. HPF is a small peptide that dimerizes 70S ribosomal subunits to form inactive 100S subunits[34, 35]. It plays an important role in stress response, nutrition limitation and protects ribosomal pools from degradation[36–38]. Previous studies in *S. aureus* have shown that SigB and CodY regulate *hpf* expression in response to heat and nutritional stress[36]. iModulon structure confirms the role of the CodY and suggests and additional layer of control by CcpA (Figure 3.4a).

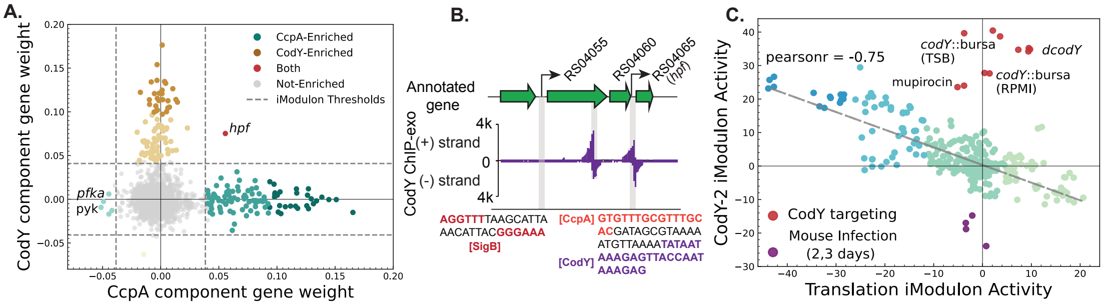


Figure 3.4: Coordination of translation with metabolism. a) Gene weights in CcpA and CodY iModulons shows that only the *hpf* is enriched in both iModulons. b) Upstream region of *hpf* gene with its two alternative transcription start sites. Two CodY binding sites were detected by ChIP-exo (purple bars). The previously recognized SigB(red) and Cody(purple) binding sites and newly proposed CcpA (orange) binding site are highlighted. c) The negative correlation between CodY and Translation iModulon suggests coordination of metabolism and translation in S. aureus.

ChIP-exo data from our previous work found two CodY binding sites in the regulatory region of the *hpf* gene (Figure 3.4b)[39]. To confirm the role of CcpA in *hpf* regulation, we searched for catabolite repressor protein motif (WTGNNARCGNWWWCAW) in the same region. A matching motif was found in the region between the two CodY binding peaks (Figure 3.4b). This architecture, with two CodY binding sites flanking the CcpA binding site, is also found in the regulatory region of *B. subtilis* BCAA operon where both regulators contribute to the expression of the operon genes[40]. The signal from expression data and the presence of binding motifs suggests that CcpA regulates *hpf* along with previously identified regulators CodY and SigB.

In addition to coordinated regulation of translation associated *hpf* gene, CodY activity was also strongly correlated with Translation iModulon activity. In contrast, CcpA and Translation iModulon activities showed little correlation between them (Figure B.4). Similar to CcpA and CodY activity correlation, *codY* knockout and stringent response activation by mupirocin also disrupted correlation with Translation iModulon (Figure 3.4c). This also suggested that the signal controlling Translation iModulon gene expression also works ‘upstream’ of CodY as interruption of CodY had little effect on Translation iModulon activity. While the coordination of the two iModulon activity is apparent, we were unable to further interrogate the nature of this relationship since the signal behind the Translation iModulon is yet to be identified.

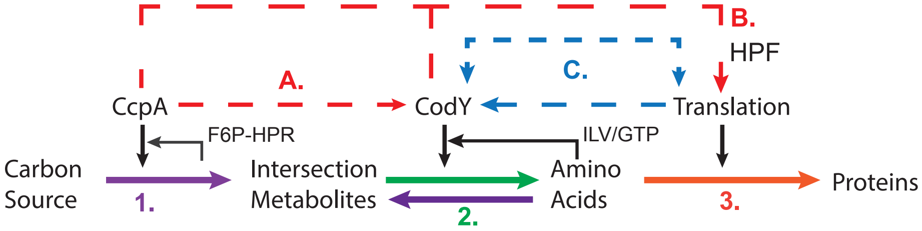


Figure 3.5: Coarse-grain model of protein synthesis in *S. aureus*. The solid lines represent the parts of the protein synthesis pathway controlled by Ccpa (purple) and CodY (green). The dashed lines represent new proposed roles of these regulators in (A) coordinating carbon and nitrogen metabolism and (B,C) linking metabolic gene expression with expression of translation associated proteins.

3.4 Discussion

Based on the data presented here, we propose a coarse grained model of transcriptional regulation of metabolism involved in protein synthesis in *S. aureus* USA300 strains (Figure 3.5). It is closely based on the model of proteome coordination in *E. coli* and extends these principles to non-model pathogenic organism[15]. The coarse grain model simplifies the metabolism underlying protein synthesis into three steps; (1) the generation of precursors from carbon sources, (2) biosynthesis of amino acids from precursors or direct transport from the medium and (3) synthesis of peptides from amino acids via translation.

The generation of precursors from carbon sources is largely regulated by CcpA (purple arrow). CcpA represses alternate carbon sources (including amino acids such as proline, glutamine and aspartate) in the presence of preferred carbon and regulates other key aspects of central metabolism such as gluconeogenesis and TCA cycle that are necessary to generate various precursors[1, 10, 22, 41, 42]. The precursors in our model are represented by the intersection metabolites derived from the USA300 GEM (Table B.1).These precursors are then converted to amino acids via CodY regulated gene products (green arrow)[39, 42, 43].

Our analysis suggests that *S. aureus* USA300 strains coordinate their CcpA and CodY activity to regulate carbon and nitrogen flow through the system. Metabolic modeling in CDMG shows increased flux through amino acid biosynthetic reactions when compared to CDM. The results of this TRN agnostic metabolic model agrees with the increased CodY activity in CDMG and other glucose containing media. Despite close coordination of metabolic flux at different intersections between CcpA and CodY reactions, it is still not clear how CcpA and CodY activities are coordinated. In E.coli, Kochanowski et al. have observed similar coordination between anabolic and catabolic fractions of metabolism[44]. The authors attribute active regulation by Crp and passive changes in metabolic fluxes in response to change in metabolite concentrations as the source of the coordination. Additionally, we also found a feed forward regulation whereby CcpA and CodY control the expression of the gene encoding HPF protein which sequesters ribosomes into inactive 100S forms, suggesting a mechanism by which translation is coordinated with metabolic state of the cell [34, 36].

Lastly, the activity of Translation iModulon is also closely correlated with CodY activity, which may act as an additional layer of coordination between metabolism and translation. However, we could not identify the signal or regulator controlling Translation activity. Ribosomal RNA (rRNA) expression is regulated by ppgpp during stringent response which can be activated by mupirocin treatment[29, 45]. We therefore expected mupirocin to also have an effect on Translation iModulon activity, but we found that while CodY activity increased in response to mupirocin as expected, there was minimal change in Translation activity (Figure 3.4c). This suggests that stringent response, at least when induced by mupirocin treatment, does not play a major role in expression of Translation iModulon genes.

The analysis of the coarse grained model of metabolic gene regulation presented here was enabled by a computable model of TRN. iModulons enable us to query the TRN at multiple-scales, giving insights into TRN from single gene membership level to global coordination of regulators. By modularizing the TRN, our analysis enabled us to unravel complex regulatory and metabolic interactions to understand regulation of central metabolism one regulator at a time. This modularization can also be used to continually expand on the presented model. For example, our previous work have shown that Translation iModulon activity in *E. coli* and *S. aureus* is closely correlated with stress associated alternate sigma factors[20, 21]. This points to a possible entry-point for coordination of general stress response with metabolism and protein synthesis. Similarly, we have also found that both PyrR and PurR activity is correlated with CodY and CcpA which may provide insights into regulation of nucleotide biosynthesis in response to carbon or nitrogen availability. While we mainly focused on 3 iModulons- CcpA, CodY and Translation- the current model contains 76 total iModulons, each of them rich with information about transcriptional regulation and physiology of *S. aureus*.

3.5 Materials and Methods

3.5.1 Strains and Growth Conditions

The *S. aureus* USA300 isolate LAC or its derivative JE2 were used to collect the new RNA sequencing data in this study. The complete description and condition for each of the samples can be found in the model sample table. For RNA sequencing from knock samples, isolated from the Nebraska Transposon Mutant Library were utilized[46]. Unless specified otherwise, samples were grown in duplicates in 20mL of respective media until they reached the O.D600nm of 0.5. 3 mL of culture was harvested and immediately mixed with 6 mL of Qiagene RNA-protect Bacteria Reagent, and incubated at room temperature for 5 minutes. The supernatant was decanted after the samples were centrifuged for 10 mins and 17,500 RPM. The remaining cell pellets were stored in -80C until they were prepared for RNA extraction.

3.5.2 RNA extraction and sequencing

Total RNA was isolated from the cell pellet in the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase treatment was performed for 30 min at room temperature. The ribosomal RNA was removed using RiboRid protocol, as described before[47]. RNA was quantified using a Nanodrop and quality assessed by running an RNA nano chip on a bioanalyser. A Swift RNA Library Kit was used following the manufacturer's protocol to create sequencing libraries.

3.5.3 Processing RNA sequencing data for iModulon calculation

The iModulons were calculated from publically available RNA sequencing data from SRA and the newly collected data in this study using pymodulon python package [48]. The steps used to calculate the iModulons described here were all completed using this package. All RNA sequencing data labeled with *S. aureus* taxonomic ID was downloaded and manually curated to obtain only the samples that were from USA300 isolates. Raw fastq files from curated samples were downloaded, trimmed with TrimGalore and were then aligned to the USA300 TCH1516 genome (NC_010079, NC_012417, NC_010063) using Bowtie2 [49]. QC/QA stats were collected on each sample using MultiQC and samples that did not pass the QC thresholds (e.g. low read depth, low correlation between replicates, missing metadata) were discarded[50]. Transcripts per million (TPM) was calculated from the remaining high quality RNA sequencing samples. TPM were log transformed and normalized to a control condition within the same BioProject.

3.5.4 Calculating iModulons from RNA sequencing data

Scipy's implementation of FastICA was applied to log transformed and normalized TPM data to generate independent components (ICs) and their activities[51, 52]. Unlike other decomposition methods, ICA requires the number of dimensions to be calculated as an input. Therefore, various models with different dimensionality were created and the one that maximized regulatory iModulons and minimized single gene iModulon was chosen[53]. The iModulons were then automatically annotated if they overlapped significantly with a curated list of known or predicted regulons and genomic features (e.g. prophages, SCCMec, ACME etc) in *S. aureus.* Other iModulons such as 'Translation' or 'Autolysin' were manually annotated as all genes contained within the iModulons have a single function.

3.5.5 Genomic scale modeling of *S. aureus* USA300 metabolism

USA300 specific Genome scale model (GEM) iYS854 was used for all metabolic simulations in the paper. Exchange rate of amino acids, glucose, ammonium and acetate were adjusted to constrain the model to CDM or CDMG specific conditions as described in detail before[30]. Briefly, the uptake or secretion rate for each metabolite from Halsey et al. were normalized by growth-rate, to get growth adjusted solute uptake rate[10]. The exchange rates were then constrained to +/-15% of uptake and exchange rate to account for variance in the data. Once constrained the model was then used to calculate flux each media using pFBA as implemented in the cobrapy package[31, 32]. To get CodY iModulon specific flux, genes in the CodY iModulon were first mapped to metabolic reactions using gene product rule (GPR). The absolute value of fluxes from the pFBA solution for the CodY reactions were then summed to get the final CodY iModulon flux. To calculate valid amino acid biosynthesis pathway specific flux distribution, the solution spaces of CDM and CDMG specific models were sampled 10,000 times using the Artificial Centering Hit-and-Run algorithm [54]. Next, the reactions in each amino acid biosynthetic pathway was determined with the MinSpan algorithm[55]. Minspan calculates the set of shortest metabolic pathways that are linearly independent of one another and span the null space of the input model. Each independent pathway defines a mass balanced set of reactions and therefore enables unbiased modularization of metabolism into biologically meaningful pathways. The sampled fluxes ($\mathbf{v}$) can therefore be represented as linear weightings ($\boldsymbol{\alpha}$) of minspan pathways ($\mathbf{P}$).

$$\mathbf{v} = \mathbf{P} \cdot \mathbf{a}$$

The sampled fluxes were converted to pathway specific weightings (pathway fluxes) using the minspan matrix. Pathways containing amino acid biosynthesis were manually curated and only amino acid biosynthesis pathways that did not appear in multiple MinSpan pathways were used for analysis as they can be easily interpreted and does not require analyzing linear combinations of multiple pathways.

Lastly, the intersection metabolites were determined by comparing all metabolites that were involved in at least one CodY and one CcpA reaction. The common metabolites ADP, ATP, CO2, coenzyme A, H2O, hydrogen atom, sodium ion, NAD, NADH, NADP, NADPH, ammonium (NH4), and phosphate were excluded from this designation.

3.5.6 Motif enrichment

The 150 base-pairs upstream of *hpf* gene (USA300HOU_RS04065) was scanned for CcpA motif (WTGNNARCGNWWWCAW) using Find Individual Motif Occurence (FIMO) within the MEME suite[56, 57].

3.6 Acknowledgements

Chapter 3, in part, is currently being prepared for submission for publication: **Poudel S**, Hefner Y, Szubin R, Sastry A, Gao Y, Nizet V, and Palsson B Ø. "Coupling of CcpA and CodY activities coordinates carbon and nitrogen metabolism associated gene expression in *S. aureus* USA300 strains." The dissertation author was the primary author.

3.7 References

1. Potter, A. D., Butrico, C. E., Ford, C. A., Curry, J. M., Trenary, I. A., Tummarakota, S. S., Hendrix, A. S., Young, J. D. & Cassat, J. E. Host nutrient milieu drives an essential role for aspartate biosynthesis during invasive Staphylococcus aureus infection. en. *Proc. Natl. Acad. Sci. U. S. A.* (May 2020).

2. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V. & Palsson, B. Ø. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

3. Diep, B. A., Stone, G. G., Basuino, L., Graber, C. J., Miller, A., des Etages, S.-A., Jones, A., Palazzolo-Ballance, A. M., Perdreau-Remington, F., Sensabaugh, G. F., DeLeo, F. R. & Chambers, H. F. The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant Staphylococcus aureus. en. *J. Infect. Dis.* **197,** 1523–1530 (June 2008).

4. Lade, H., Park, J. H., Chung, S. H., Kim, I. H., Kim, J.-M., Joo, H.-S. & Kim, J.-S. Biofilm Formation by Staphylococcus aureus Clinical Isolates is Differentially Affected by Glucose and Sodium Chloride Supplemented Culture Media. en. *J. Clin. Med. Res.* **8** (Nov. 2019).

5. DeMars, Z. & Bose, J. L. Redirection of Metabolism in Response to Fatty Acid Kinase in Staphylococcus aureus. en. *J. Bacteriol.* **200** (Oct. 2018).

6. Melter, O. & Radojevič, B. Small colony variants of Staphylococcus aureus–review. en. *Folia Microbiol.* **55,** 548–558 (Nov. 2010).

7. Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C. & Collins, J. J. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. en. *Cell* **177,** 1649–1661.e9 (May 2019).

8. Lopatkin, A. J., Stokes, J. M., Zheng, E. J., Yang, J. H., Takahashi, M. K., You, L. & Collins, J. J. Bacterial metabolic state more accurately predicts antibiotic lethality than growth rate. en. *Nat Microbiol* (Aug. 2019).

9. Gaupp, R., Lei, S., Reed, J. M., Peisker, H., Boyle-Vavra, S., Bayer, A. S., Bischoff, M., Herrmann, M., Daum, R. S., Powers, R. & Somerville, G. A. Staphylococcus aureus metabolic adaptations during the transition from a daptomycin susceptibility phenotype to a daptomycin nonsusceptibility phenotype. en. *Antimicrob. Agents Chemother.* **59,** 4226–4238 (July 2015).

10. Halsey, C. R., Lei, S., Wax, J. K., Lehman, M. K., Nuxoll, A. S., Steinke, L., Sadykov, M., Powers, R. & Fey, P. D. Amino acid catabolism in Staphylococcus aureus and the function of carbon catabolite repression. *MBio* **8** (2017).

11. Ding, Y., Liu, X., Chen, F., Di, H., Xu, B., Zhou, L., Deng, X., Wu, M., Yang, C.-G. & Lan, L. Metabolic sensor governing bacterial virulence in Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E4981–90 (Nov. 2014).

12. Thomas, V. C., Sadykov, M. R., Chaudhari, S. S., Jones, J., Endres, J. L., Widhelm, T. J., Ahn, J.-S., Jawa, R. S., Zimmerman, M. C. & Bayles, K. W. A central role for carbon-overflow pathways in the modulation of bacterial cell death. en. *PLoS Pathog.* **10,** e1004205 (June 2014).

13. Orth, J. D., Thiele, I. & Palsson, B. Ø. *What is flux balance analysis?* 2010.

14. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. en. *Nat. Rev. Microbiol.* **10,** 291–305 (Feb. 2012).

15. You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C. W., Wang, Y.-P., Lenz, P., Yan, D. & Hwa, T. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. en. *Nature* **500,** 301–306 (Aug. 2013).

16. Basan, M., Honda, T., Christodoulou, D., Hörl, M., Chang, Y.-F., Leoncini, E., Mukherjee, A., Okano, H., Taylor, B. R., Silverman, J. M., Sanchez, C., Williamson, J. R., Paulsson, J., Hwa, T. & Sauer, U. A universal trade-off between growth and lag in fluctuating environments. en. *Nature* **584,** 470–474 (Aug. 2020).

17. Erickson, D. W., Schink, S. J., Patsalo, V., Williamson, J. R., Gerland, U. & Hwa, T. A global resource allocation strategy governs growth transition kinetics of Escherichia coli. en. *Nature* **551,** 119–123 (Nov. 2017).

18. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336,** 601–604 (May 2012).

19. Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M. & Palsson, B. O. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. en. *Cell Syst* **2,** 260–271 (Apr. 2016).

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

21. 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. Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. en. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 17228–17239 (July 2020).

22. Seidl, K., Müller, S., François, P., Kriebitzsch, C., Schrenzel, J., Engelmann, S., Bischoff, M. & Berger-Bächi, B. Effect of a glucose impulse on the CcpA regulon in Staphylococcus aureus. en. *BMC Microbiol.* **9,** 95 (May 2009).

23. Rychel, K., Sastry, A. V. & Palsson, B. O. Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. en. *Nat. Commun.* **11,** 6338 (Dec. 2020).

24. Yoo, R., Rychel, K., Poudel, S., Al-bulushi, T., Yuan, Y., Chauhan, S., Lamoureux, C., Palsson, B. O. & Sastry, A. *Machine learning of all Mycobacterium tuberculosis H37Rv RNA-seq data reveals a structured interplay between metabolism, stress response, and infection* en. July 2021.

25. Chauhan, S. M., Poudel, S., Rychel, K., Lamoureux, C., Yoo, R., Al Bulushi, T., Yuan, Y., Palsson, B. O. & Sastry, A. V. Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile Sulfolobus acidocaldarius. en. *Front. Microbiol.* **12,** 753521 (Oct. 2021).

26. Yuan, Y., Seif, Y., Rychel, K., Yoo, R., Chauhan, S., Poudel, S., Al-bulushi, T., Palsson, B. O. & Sastry, A. *Pan-genomic analysis of transcriptional modules across Salmonella Typhimurium reveals the regulatory landscape of different strains* en. Jan. 2022.

27. Kaiser, J. C., King, A. N., Grigg, J. C., Sheldon, J. R., Edgell, D. R., Murphy, M. E. P., Brinsmade, S. R. & Heinrichs, D. E. Repression of branched-chain amino acid synthesis in Staphylococcus aureus is mediated by isoleucine via CodY, and by a leucine-rich attenuator peptide. en. *PLoS Genet.* **14,** e1007159 (Jan. 2018).

28. Radin, J. N., Kelliher, J. L., Párraga Solórzano, P. K. & Kehl-Fie, T. E. The Two-Component System ArlRS and Alterations in Metabolism Enable Staphylococcus aureus to Resist Calprotectin-Induced Manganese Starvation. en. *PLoS Pathog.* **12,** e1006040 (Nov. 2016).

29. Reiss, S., Pané-Farré, J., Fuchs, S., François, P., Liebeke, M., Schrenzel, J., Lindequist, U., Lalk, M., Wolz, C., Hecker, M. & Engelmann, S. Global analysis of the Staphylococcus aureus response to mupirocin. en. *Antimicrob. Agents Chemother.* **56,** 787–804 (Feb. 2012).

30. Seif, Y., Monk, J. M., Mih, N., Tsunemoto, H., Poudel, S., Zuniga, C., Broddrick, J., Zengler, K. & Palsson, B. O. A computational knowledge-base elucidates the response of Staphylococcus aureus to different media types. en. *PLoS Comput. Biol.* **15,** e1006644 (Jan. 2019).

31. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. en. *BMC Syst. Biol.* **7,** 74 (Aug. 2013).

32. Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., *et al.* Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6,** 390 (2010).

33. Proctor, R. A. Role of folate antagonists in the treatment of methicillin-resistant Staphylococcus aureus infection. en. *Clin. Infect. Dis.* **46,** 584–593 (Feb. 2008).

34. Puri, P., Eckhardt, T. H., Franken, L. E., Fusetti, F., Stuart, M. C. A., Boekema, E. J., Kuipers, O. P., Kok, J. & Poolman, B. Lactococcus lactis YfiA is necessary and sufficient for ribosome dimerization. en. *Mol. Microbiol.* **91,** 394–407 (Jan. 2014).

35. Gohara, D. W. & Yap, M.-N. F. Survival of the drowsiest: the hibernating 100S ribosome in bacterial stress management. en. *Curr. Genet.* **64,** 753–760 (Aug. 2018).

36. Basu, A., Shields, K. E., Eickhoff, C. S., Hoft, D. F. & Yap, M.-N. F. Thermal and nutritional regulation of ribosome hibernation in Staphylococcus aureus. en. *J. Bacteriol.* (Oct. 2018).

37. Basu, A. & Yap, M.-N. F. Ribosome hibernation factor promotes Staphylococcal survival and differentially represses translation. en. *Nucleic Acids Res.* **44,** 4881–4893 (June 2016).

38. Lipońska, A. & Yap, M.-N. F. Hibernation-Promoting Factor Sequesters Staphylococcus aureus Ribosomes to Antagonize RNase R-Mediated Nucleolytic Degradation. en. *MBio* **12,** e0033421 (Aug. 2021).

39. Gao, Y., Poudel, S., Seif, Y., Shen, Z. & Palsson, B. O. *Elucidating the CodY regulon in Staphylococcus aureus USA300 substrains* en. Jan. 2021.

40. Fujita Yasutaro, Satomura Takenori, Tojo Shigeo & Hirooka Kazutake. CcpA-Mediated Catabolite Activation of the Bacillus subtilis ilv-leu Operon and Its Negation by Either CodY- or TnrA-Mediated Negative Regulation. *J. Bacteriol.* **196,** 3793–3806 (Nov. 2014).

41. Nuxoll, A. S., Halouska, S. M., Sadykov, M. R., Hanke, M. L., Bayles, K. W., Kielian, T., Powers, R. & Fey, P. D. CcpA regulates arginine biosynthesis in Staphylococcus aureus through repression of proline catabolism. en. *PLoS Pathog.* **8,** e1003033 (Nov. 2012).

42. Sonenshein, A. L. Control of key metabolic intersections in Bacillus subtilis. en. *Nat. Rev. Microbiol.* **5,** 917–927 (Dec. 2007).

43. Waters, N. R., Samuels, D. J., Behera, R. K., Livny, J., Rhee, K. Y., Sadykov, M. R. & Brinsmade, S. R. A spectrum of CodY activities drives metabolic reorganization and virulence gene expression in *Staphylococcus aureus. Mol. Microbiol.* **101,** 495–514 (Aug. 2016).

44. Kochanowski, K., Okano, H., Patsalo, V., Williamson, J., Sauer, U. & Hwa, T. Global coordination of metabolic pathways in Escherichia coli by active and passive regulation. en. *Mol. Syst. Biol.* **17,** e10064 (Apr. 2021).

45. Samarrai, W., Liu, D. X., White, A.-M., Studamire, B., Edelstein, J., Srivastava, A., Widom, R. L. & Rudner, R. Differential responses of Bacillus subtilis rRNA promoters to nutritional stress. en. *J. Bacteriol.* **193,** 723–733 (Feb. 2011).

46. Fey, P. D., Endres, J. L., Yajjala, V. K., Widhelm, T. J., Boissy, R. J., Bose, J. L. & Bayles, K. W. A genetic resource for rapid and comprehensive phenotype screening of nonessential Staphylococcus aureus genes. en. *MBio* **4,** e00537–12 (Feb. 2013).

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

48. Sastry, A. V., Poudel, S., Rychel, K., Yoo, R., Lamoureux, C. R., Chauhan, S., Haiman, Z. B., Al Bulushi, T., Seif, Y. & Palsson, B. O. *Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks* en. July 2021.

49. Highlander, S. K., Hulten, K. G., Qin, X., Jiang, H., Yerrapragada, S., Mason, E. O., Shang, Y., Williams, T. M., Fortunov, R. M., Liu, Y., Igboeli, O., Petrosino, J., Tirumalai, M., Uzman, A., Fox, G. E., Cardenas, A. M., Muzny, D. M., Hemphill, L., Ding, Y., Dugan,

S., Blyth, P. R., Buhay, C. J., Dinh, H. H., Hawes, A. C., Holder, M., Kovar, C. L., Lee, S. L., Liu, W., Nazareth, L. V., Wang, Q., Zhou, J., Kaplan, S. L. & Weinstock, G. M. Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. *BMC Microbiol.* **7,** 99 (Nov. 2007).

50. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. en. *Bioinformatics* **32,** 3047–3048 (Oct. 2016).

51. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. en. *Neural Netw.* **13,** 411–430 (May 2000).

52. Eric Jones, Travis Oliphant, Pearu Peterson and others. *SciPy: Open Source Scientific Tools for Python* 2001.

53. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O. & Sastry, A. V. Optimal dimensionality selection for independent component analysis of transcriptomic data. en. *BMC Bioinformatics* **22,** 584 (Dec. 2021).

54. Megchelenbrink, W., Huynen, M. & Marchiori, E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. en. *PLoS One* **9,** e86587 (Feb. 2014).

55. Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., Schellenberger, J. & Palsson, B. O. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. en. *Mol. Syst. Biol.* **10,** 737 (July 2014).

56. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. en. *Nucleic Acids Res.* **37,** W202–8 (July 2009).

57. Schumacher, M. A., Sprehe, M., Bartholomae, M., Hillen, W. & Brennan, R. G. Structures of carbon catabolite protein A-(HPr-Ser46-P) bound to diverse catabolite response element sites reveal the basis for high-affinity binding to degenerate DNA operators. en. *Nucleic Acids Res.* **39,** 2931–2942 (Apr. 2011).

Chapter 4

Interpreting roles of mutations in the emergence of *S. aureus* USA300 strains with genetics and independent component analysis of gene expression

4.1 Abstract

The *Staphylococcus aureus* clonal complex 8 (CC8) can be divided into several subtypes containing one of community associated methicillin resistant *S. aureus* (CA-MRSA) USA300, hospital-associated MRSA (HA-MRSA) USA500 or basal methicillin susceptible *S.*

aureus (MSSA) strains. This makes CC8 an ideal clade to study the emergence of mutations important for resistance and community spread. Gene level analysis comparing USA300 against MSSA and HA-MRSA strains have revealed key horizontally acquired genes important for its rapid spread in the community. However, efforts to define contributions of point mutations and indels have been confounded by strong linkage disequilibrium resulting from clonal propagation. To break down this confounding effect, we combined genetic association testing with a model of transcriptional regulatory network (TRN) to find candidate mutations that led to changes in gene regulations. We used a De Bruijn graph genome-wide association study (DBGWAS) to enrich mutations unique to the USA300 lineages. Next, we modeled the TRN by using Independent Component Analysis on 628 RNA sequencing samples from USA300 and non-USA300 CC8 strains. Our models predicted several genes with strain-specific altered expression patterns as well as DBGWAS enriched mutations. Examination of the regulatory region of one of the genes that were enriched by both approaches, *isdH*, revealed a 38 base pair deletion containing Fur binding site and a conserved SNP which likely led to the altered expression levels. Our results demonstrate the utility of modeling gene regulation as a promising method to address the limits of genetic approaches when studying emerging pathogenic strains.

4.2 Introduction

Comparative genomic methods are an important tool in understanding the emergence and evolution of new strains of pathogens. In *S. aureus* alone, whole genome comparisons have enabled rapid characterization of genetic basis for antibiotic resistance, increased virulence, host specificity and altered metabolic capabilities [1–5]. However, genome-wide linkage disequilibrium and strong lineage structuring currently limits the differentiation of causative alleles from genetically linked ones. By calculating lineage level associations, methods like bugwas address these issues for single, recurring phenotypes like antibiotic resistance[6]. Endemic strains, on the other hand, exhibit multiple complex phenotypes that may contribute to their emergence and proliferation. For example, USA300 strains carry antibiotic resistance cassettes, Panton Valentine Leukocidin (PVL) associated with pyomastitis, increased ability to colonize locations outside of the nasopharynx, etc. As these strains often emerge clonally from closely related 'basal strains,' efforts to discern causal mutations that lead to their increased clinical burden is hampered by strong population-stratification and genome-wide linkage disequilibrium[7–9]. Though recombination at species level is common in *S.aureus*, within clade recombination rates tend to be lower, thus preserving the linkage disequilibrium[8, 10–12]. Due to this limitation, studies of emerging strains often focus on gene level analysis such as acquisition of mobile genetic elements or loss of gene function while determining the possible phenotypic effect (if any) of all enriched Single Nucleotide Polymorphisms (SNPs) remains challenging[13].

Even if experimentally intractable, the large possible phenotypic space of an organism can be explored quickly with computational models. Combined with GWAS, computational modeling can be used as a sieve to filter enriched mutations with potential phenotypic effects and therefore find candidate causal mutations[14–16]. Here, we used De Bruijn graph GWAS (DBGWAS) to enrich mutations associated with the endemic USA300 strain within clonal complex 8 (CC8)[17]. Due to clonal expansion of USA300 strains from their progenitors within CC8, the enriched USA300 specific mutations were in high linkage disequilibrium. Further complicating the matter, we found that almost all mutations enriched within ORFs were unique to USA300 lineage and not found in any other clonal complexes, precluding identification of potential causative mutations by homoplasy.

To get around these limitations of genetic approach, we built an ICA based model of transcription regulation using 628 publicly available RNA sequencing samples from CC8 strains. By factoring the RNA sequencing data into a series of signals and their activities, the ICA model of TRN shows both the static gene-regulator interaction and the dynamic activity of these interactions in a sample specific manner[18]. However, ICA is a generalized signal extraction algorithm and therefore does not distinguish between biological sources of signals like regulatory elements and 'artificial' sources that can be created by sourcing data from multiple strains. Therefore, in addition to signals associated with gene regulators, ICA also outputs signals associated with strain-specific changes in the gene regulation. By utilizing RNA sequencing data from hundreds of samples to extract genes with strain-specific expression patterns, this modeling approach is more likely to find strain-specific differences than previous approaches that focus on specific conditions [19, 20]. The model revealed several genes with distinct expression patterns in USA300 strains that were also associated with a DBGWAS enriched mutation. Close analysis of one of these genes, *isdH*, which encodes a haptoglobin binding protein, showed several mutations in the gene regulatory region including deletion of the transcription factor Fur binding site in the USA300 strain. Overall, our analysis shows how models of TRN can be used to extend the limits of current GWAS approaches when studying emerging and endemic populations of bacterial pathogens.

4.3 Results

4.3.1 Classifying USA300 and non-USA300 genomes based on genetic markers

We sought to compare the genetic differences between USA300 CA-MRSA strains and other subtypes within CC8 that have lower clinical and community burden. Given that both subtypes exist within the same clonal complex, this comparison allowed us to probe the genetic basis for the success of USA300 strains with limited confounding effects of different genetic backgrounds. We analyzed 2038 *S.aureus* CC8 genomes which formed a closed pangenome, suggesting that the sampled genomes mostly captured the gene level variations within the clonal complex (Figure C.1a). The CC8 pangenome consisted of 19176 unique genes with 2291 core genes that were present in at least 95% of the genomes analyzed. Among the remainder of the genes, 931 were categorized as accessory genes and 15954 were uniquely found in less than 5% of the genomes. Interestingly, we found a larger number of unique alleles in the ORFs than in proximal 3' and 5' regions, indicating the presence of greater genetic variation among ORFs than in the neighboring regulatory regions (Figure 4.1a).

Next, we classified the CC8 genomes into USA300 and non-USA300 strains using Genetic Marker Inference (GMI). GMI was previously developed to rapidly and systematically identify different subclades within inner-CC8 strictly based on genetic markers [21]. In this scheme, USA300 genomes can be differentiated from non-USA300 genomes by the presence of either SCCMecIVa or the presence of Panton-Valentine Leukocidin (PVL) in case of methicillin sensitive *S.aureus* (MSSA). We added additional criteria that all genomes identified as USA300 by GMI form a distinct subclade before they are labeled as USA300 i.e. PVL or SCCMECIVa positive genomes that grouped separately from other USA300 strains in the phylogenetic tree were not labeled as USA300. Using the SCCMecFinder tool, we detected SCCMec cassettes in 1588 genomes of which 1358 were SCCMecIVa positive[22]. We also found 1431 PVL positive genomes using BLASTn search with PVL encoding genes from USA300 TCH1516 (USA300HOU_RS07645, USA300HOU_RS07650) as reference. Lastly, we reconstructed the CC8 phylogenetic tree based on core Single Nucleotide Polymorphisms (SNPs) and rooted the tree using strain D592 (CC5) as an outgroup (Figure 4.1b).

To identify the root of the USA300 clades, we first traversed up nodes of the phylogenetic tree starting from known USA300 strain TCH1516 and determined the number of strains, fraction PVL positive and fraction SCCMecIVa positive for each node during traversal. The root was placed at the last node where greater than 90% of the strains within the subclade represented by the node were SCCMecIVa and PVL positive (Figure 4.1b). As phylogenetic trees are nested, root finding with this procedure is not dependent on the starting USA300 strain. Same root was identified when the procedure was initialized with another well known USA300 reference strain FPR3757 (Figure C.1b). Combining the genetic markers with phylogenetic grouping led to the classification of 1449 genomes as USA300 and 589 genomes as non-USA300 (Figure 4.1c). Strains previously identified as 'early USA300' were not part of our USA300 classification[21]. While many of these strains are PVL positive, they have variable SCCMec types and therefore are likely to be genetically distinct from the endemic USA300 strains.

4.3.2 Enriching USA300 specific genes and mutations using DBGWAS

After classifying the genomes into USA300 and non-USA300 strains, we identified genes and mutations associated with each subtype by using the De Bruijn graph Genome Wide Associ-

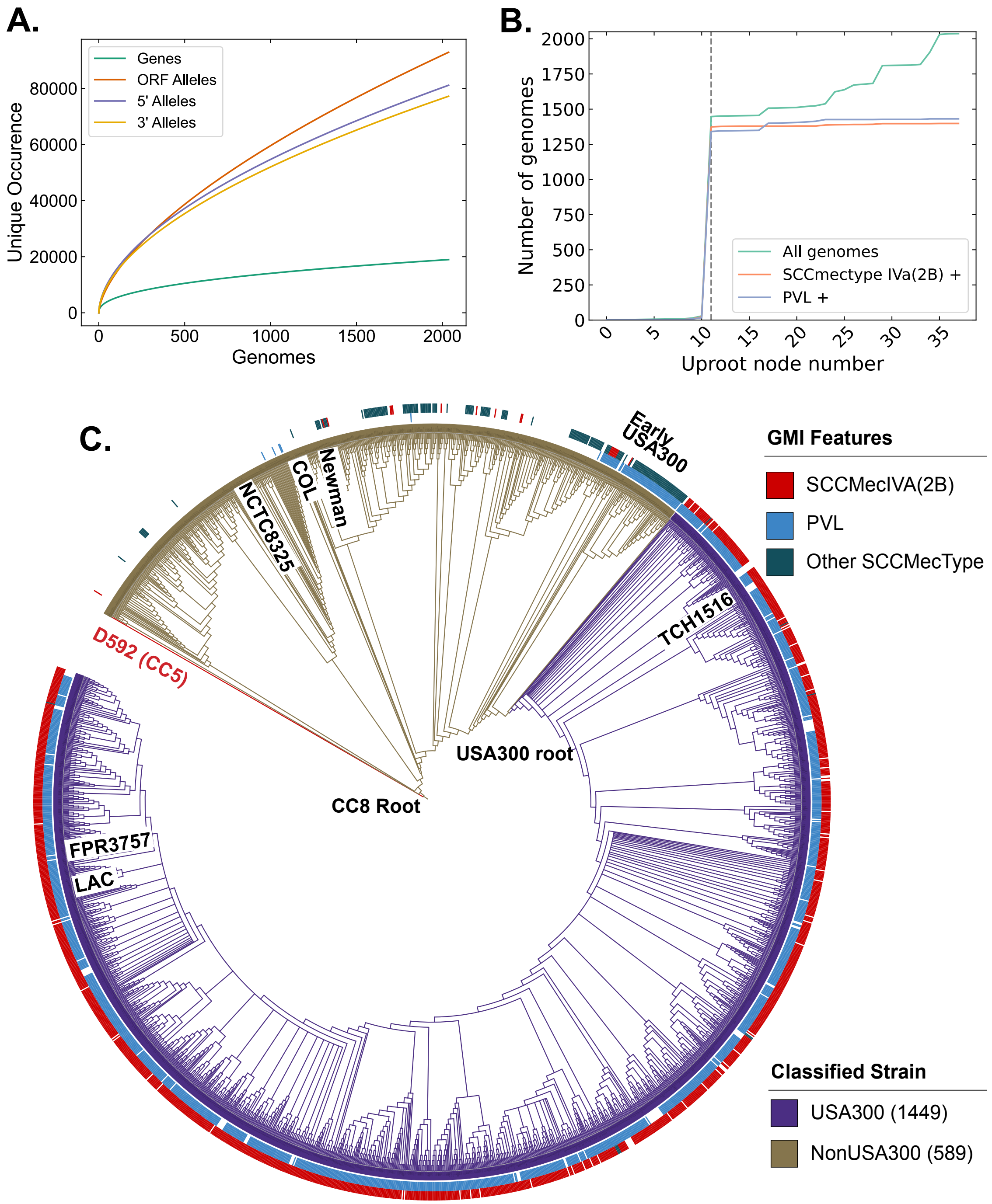


Figure 4.1: CC8 pangenome and phylogeny. (a) Pangenomic analysis of CC8 genomes shows the distribution of genes and mutations in ORFs and regulatory regions. (b) Prevalence of USA300 specific genetic markers, PVL and SCCMecIVa, as you traverse up the phylogenetic tree from TCH1516. The gray dashed line represents the node where the USA300 root is placed. (c) Phylogenetic tree of CC8 genomes classified into USA300 and non-USA300 strains.

ation Study (DBGWAS)[17]. DBGWAS provides a reference-genome free method for conducting GWAS analysis in prokaryotes by building a compacted De Brujin Graph to represent the pan genome of input sequences. The nodes of the graph represent unique compacted k-mers that are joined by edges to other nodes with k-mers that appear adjacent to it in genomes. The procedure then searches for k-mers that appear with different frequencies in each classification and outputs the enriched k-mer as well as it's genetic neighborhood (called 'components') from the De Bruijn graph. Visualizing the components associated with the enriched kmers makes it easier to interpret the k-mers and makes it easy to identify large structural variations (e.g. cassette acquisition) which are often represented by multiple enriched k-mers that fall within the same component.

Many of the components were associated with genes and genetic elements expected to be enriched with USA300 strains- SCCMecIVA (the GMI marker), Arginine Catabolite Repressor Element (ACME), cap5E point mutation, multiple prophages etc were also enriched by DBGWAS. In total, we found 147 components that were enriched in this analysis, pointing to a large array of mutations that are unique to the USA300 lineage (Figure 4.2a).

Currently, DBGWAS outputs the graph consisting of the nodes with the enriched k-mers and its genetic neighborhood, but does not automatically yield the exact mutation associated with each of the significant nodes. By analyzing the structure of the component graphs with networkX, we were able to extract the exact genetic changes represented by these components[23]. Mobile genetic elements (MGEs) and large indels can be identified by a series of nodes that are all enriched in either USA300 or non-USA300 genomes (Figure 4.2b). The enrichment of multiple sequential k-mers in only one of the groups implies deletion of the sequence (or conversely insertion) in the other group. SNPs and indels smaller than the kmer-size on the other hand form

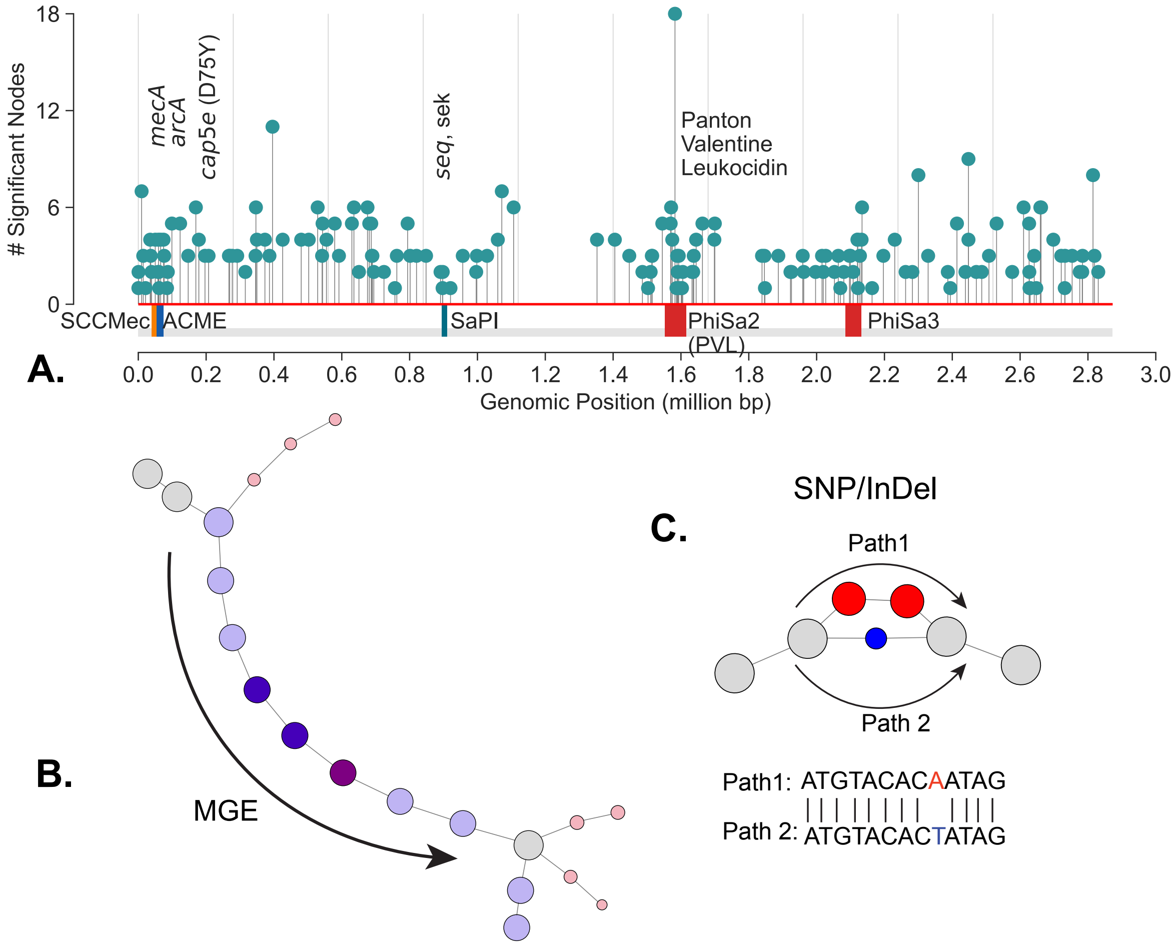


Figure 4.2: USA300 strains associated mutations. (a) DBGWAS recovers components associated with USA300 previously described markers of USA300 strains including mecA (SCCMec IVa), arcA (ACME), cap5e mutation, seq, sek and Phi-PVL. In addition, components with many other mutations scattered throughout the genome are also enriched. (b) Example of components associated with MGEs; components have an series of nodes that are enriched in one group (blue circles). (c) Example of components associated with SNP. Component graph contains a cycle around the mutation location with the paths from the cycle forming a sequence unique to either case or control group. Aligning the sequences reveals the enriched mutation.

'cycles' containing significant nodes (Figure 4.2c). Consequently, the k-mers in the nodes of each of the 'paths' around the cycle represent sequences unique to either case or control group. The enriched mutation can therefore be extracted by comparing the sequences with global alignment. Lastly, the unique sequences from each path can also be mapped to reference genomes if needed.

The exact mutations used in the subsequent sections were extracted from the components using this method.

4.3.3 Genome-wide linkage and *de novo* mutations obfuscate identification of causal mutations

Though these mutations were enriched in USA300 strains with DBGWAS, we could not attribute the prevalence of any particular mutation to selection due to strong genome-wide linkage. We quantified the linkage disequilibrium by calculating the square of the correlation coefficient ($r2$) for each of the enriched k-mer not associated with MGEs. High correlation coefficient indicates tight co-occurrence of kmers in the genomes and therefore high linkage disequilibrium between the sequences. There was a strong linkage between the k-mers that were enriched in USA300 strains. Surprisingly, even k-mers that were 1.4 million base pairs away (the maximum distance between two sites in the circular 2.8 million base pairs long *S.aureus* genome) still had $r2 > 0.9$(Figure 4.3a).

To differentiate potential causal mutations from genetically linked alleles, we searched for mutation hotspots by comparing the positions of USA300 mutations in open reading frames (ORFs) to mutations in other clonal complexes. Barring recombination events, presence of mutation hotspots in the same position in multiple clades could point to selection acting on the sequence. Therefore, we searched for prevalence of enriched mutations in other non-CC8 clades. We identified 61 SNPs within open reading frames (ORFs) that were enriched in USA300 strains. To identify mutational hotspots in other clades, we downloaded all the amino acid sequences belonging to the PATRIC genus protein family of each of the gene products encoded by the selected ORFs. [24]. The PATRIC local protein family consists of sequences of homologous

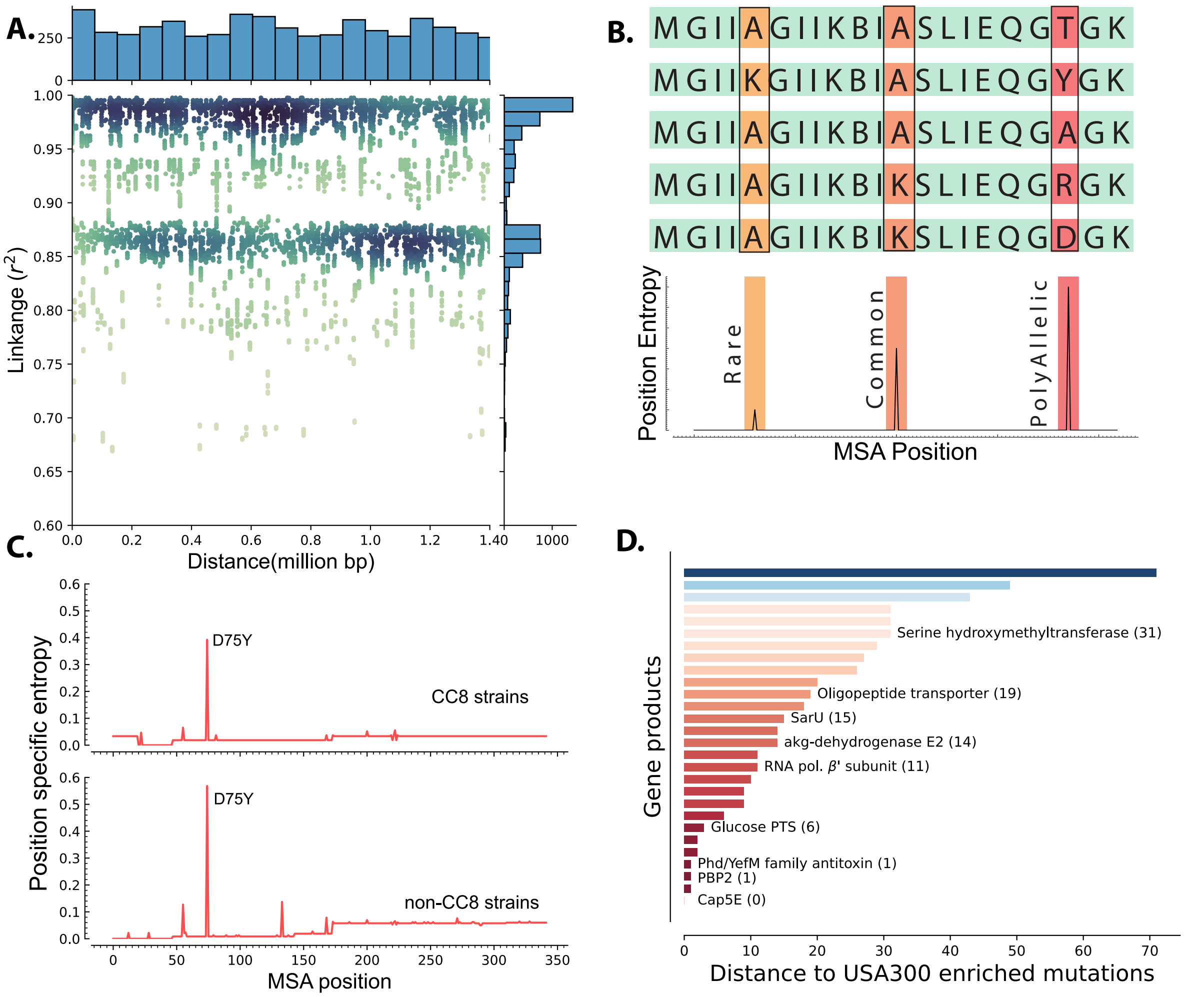


Figure 4.3: Linkage Disequilibrium and de novo mutations in USA300 strains. (a) Enriched k-mers showed high linkage disequilibrium, with some k-mers at 1.4 Mbp distance still having r2 of greater than 0.98. (b) Schematic of position specific entropy analysis. Positions with heterogeneous sequences have higher calculated entropy than more conserved sequences with fewer mutations. (c) Using position specific entropy, we only found one example of shared enriched mutation in ORFs of USA300 and non-USA300 strains. (d) Distance between the position of enriched mutation in USA300 strains and the position of the nearest entropy peak in other non-CC8 strains.

proteins within the same genus which were further filtered down to *S.aureus* species specific sequences. After filtering, each protein family comprised 2,000 to 16,000 unique sequences and the strains from which the amino acid sequences were derived spanned dozens of clades allowing for broad comparisons (Figure C.2). Lastly, we removed sequences associated with ST239 as it is thought to have emerged from large-scale recombination of ST8 and ST30 strains[25].

We determined mutation hotspots by calculating position specific allelic entropy. Allelic entropy at a given amino acid position is a function of the number of unique amino acids found in that position and the frequency of the mutation[26]. Positions where all queried sequences have the same amino acid have low entropy, while positions that have frequent amino acid substitutions (hotspots) have high entropy (Figure 4.3b). This measure allows us to quickly determine the positions of mutation hotspots while accounting for multiple possible amino acid substitutions and rare mutations. Before calculating the position-specific entropy, all sequences within each of the PATRIC local protein family were aligned with Multiple Sequence Alignment (MSA). This alignment ensures proper comparison of amino acids even when there are deletions or insertions in some of the genes in the family.

Of the 36 enriched ORF mutations only the Asp75Tyr mutation in the cap5E gene, which was previously shown to ablate capsule production in USA300 strains, was found in other strains (Figure 4.3c)[27]. Peaks in entropy corresponding to this mutation position were present in both the CC8 and non-CC8 strains while all other mutation positions were unique to CC8. Despite not having any perfect matches outside of the cap5E mutation, we found that for 28 of the mutations, a peak was present in sequences from other clades within 71 MSA positions. Together, our data suggests that mutations within ORFs in USA300 strains are likely *de novo* mutations and are not acquired through horizontal gene transfer though many of these mutations have occurred in hotspot regions (Figure 4.3d).

4.3.4 iModulon model of CC8 TRN points to mutations associated with differential regulation

The presence of genome-wide linkage and *de novo* mutations in ORFs severely limited the ability to distinguish causal SNPs contributing to increased pathogenesis in USA300 strains. The effect of some mutations, especially in ORFs, has been successfully linked to distinct phenotypes such as the absence of a capsule in USA300 and USA500 strains[27]. However, the effect of mutations associated with changes in gene regulations can be much more difficult to assess[13]. To look for mutations that may be associated with changes in transcriptional regulation, we used ICA to model gene-regulation in CC8 strains which can predict strain specific differences in expression patterns. We collected USA300 and NCTC8325 (including derivatives such as HG001) RNA-sequencing data from Sequence Read Archive (SRA). After stringent QC/QA and curation, 285 NCTC825 and 343 USA300 samples were used to create a single model of transcription regulation using ICA[18, 28]. ICA calculates independently modulated sets of genes, iModulons, and the activities of those gene sets in each sample. iModulons calculated by ICA represent distinct sources of signals in the RNA-sequencing data. While most of the signals can be associated with different regulatory elements, iModulons associated with other biological features such as mobile genetic elements, genetic backgrounds are also enriched. In Escherichia coli and Salmonella enterica Typhimurium, multi-strain ICA has been used to calculate strain-specific iModulons that represent differences in gene expression[18, 29].

In our model, two iModulons captured a large number of genes with different expression levels in the non-USA300 NCTC8325 and USA300 strains (Figure 4.4a; Figure C.3a). Most of the genes in the strain-specific iModulons belonged to mobile elements associated with USA300 strains such as ACME, SCCMec, Phi-PVL etc(Figure 4.4b; Figure C.3b). However, the iModulons also contained core genes that are present in both strains, pointing to possible differences in gene regulation.

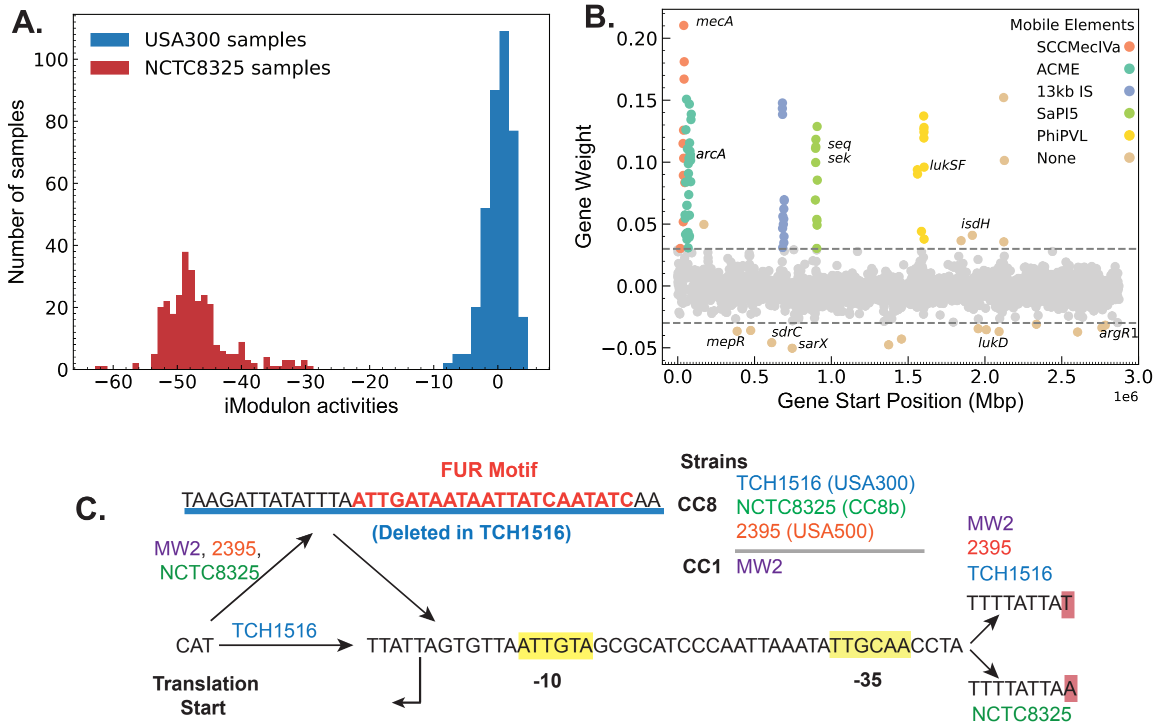


Figure 4.4: Strain-specific regulatory changes in CC8 clade. (a) ICA analysis of USA300 and NCTC8325 RNA-sequencing data identified an iModulon with strain specific activity.(b) The strain-specific iModulon contained various horizontally acquired elements (e.g. ACME, PhiPVL) that are prevalent in USA300 lineage as well as conserved genes with strain-specific expression patterns. (c) Comparing the 5' regulatory region of the gene *isdH* from various *S. aureus* strains revealed a unique deletion containing Fur binding site in USA300 reference strain TCH1516.

We mapped the enriched mutations from DBGWAS onto the core genes enriched in the strain-specific iModulon. 3 genes with mutations in the ORF or in the regulatory region were also enriched in the iModulon. Of the these genes, gene *isdH*, encoding a heme scavenger molecule showed distinct strain-specific expression levels in the 628 total RNA-sequencing profiles. K-mers that are mapped to the upstream regulatory region of the *isdH* gene were enriched by DBGWAS. Therefore, we compared the upstream regulatory region of several reference strains including TCH1516 (USA300), NCTC8325 (CC8b), 2395 (USA500). Additionally, we included MW2 (CC1

CA-MRSA) as the transcription start site (TSS) in the region has been experimentally confirmed in this strain[30]. Comparisons showed a 38 base-pair deletion in the 5' untranslated region containing a transcription factor Fur binding site (q-val=0.033e-4). This deletion was detected in all of the 1385 USA300 genomes, but only present in 95 of the 589 non-USA300 genomes. As Fur is a repressor that blocks expression in presence of iron concentration, this deletion in the Fur binding site may be responsible for the general increase in *isdH* expression observed in USA300 samples (Figure C.4a). We also found a second mutation upstream of the predicted -35 binding site that was also enriched in USA300. Interestingly, while the MW2 strain did not have the 38 bp deletion, it contained the exact upstream A to T mutation. All other base-pairs in the region were perfectly matched in between all the reference genomes. The combination of evidence from genetic and transcriptomic analysis suggests that regulation of *isdH* is therefore altered in USA300 strains compared to its non-USA300 progenitors.

4.4 Discussion

Emergence of CA-MRSA USA300 strains from HA-MRSA USA500 progenitors presents a natural experiment to probe the genetic basis for success of the USA300 lineage. However, in studying these groups, genetic methods like GWAS were limited in finding causal mutations due to genome-wide linkage disequilibrium and presence of an unexpectedly large number of *de novo* mutations unique to the USA300 lineage. Here, we demonstrated how a model of transcriptional regulation with iModulons can be used to break through the impasse created by the high linkage disequilibrium and predict candidate causal mutations. From the combined RNA sequencing dataset of USA300 and non-USA300 strains, ICA calculated iModulons that captured strain specific variation in gene expression. As expected, most genes in the iModulons were part of mobile genetic elements such as ACME and SCCMec because they have zero expression level in non-USA300 samples. However, the iModulon also contained several core genes that are present in both groups but are differentially regulated. A deeper analysis of the regulatory region of one of these genes with enriched mutation, *isdH*, revealed a deletion of a DNA segment containing the binding site of the Fur repressor. In congruence with this observation, we also found that USA300 strains with the deleted Fur binding site showed general increase in *isdH* expression level. Combining GWAS with large-scale transcriptomic modeling was therefore able to predict potential causal mutations that led to the increased clinical burden of the USA300 lineage.

While the current analysis utilized the available DNA and RNA sequencing data, the methods used here are scalable to the rapidly growing number of data in the public repositories. Indeed, with the greater scale, we can get more granular insight into subclade specific differences. The transcriptomic analysis consisted of samples primarily from the USA300 (CC8e and CC8f) clades and the CC8b clade represented by NCTC8325 and its derivatives. However, the CC8b clade is currently undersampled due to its minimal clinical burden compared to USA300. We therefore combined strains from all non-USA300 clades into a single group for GWAS. The misalignment of RNA sequencing samples from GWAS samples may explain the low number of hits that were enriched by both methods when many other unique gene expression patterns have been observed in USA300 strains. The scalability of the methods used herein will enable granular and more in-depth analysis as these sequencing databases expand rapidly.

With time, the scaling of databases may be able to resolve the issue of imbalanced sampling. On the other hand, resolving the confounding effect of linkage disequilibrium inherent in emerging and endemic strains will require a new generation of modeling methods[9]. Our current approach focuses on modeling the changes in gene regulation at the transcriptional level, but causal mutations can have any number of effects on the phenotype of the organism. New modeling methods that can systematically predict these other phenotypes are now rapidly emerging. Our recent work with Mycobacterium tuberculosis utilized a metabolic allele classifier (MACs) which combines genome scale metabolic models with machine learning to estimate biochemical effects of alleles thus mapping mutations to changes in metabolic fluxes[16]. Similarly, advances in protein structure prediction with AlphaFold2 and RosettaFold puts us at the cusp of being able to predict the effects of mutations on protein folding[31, 32]. Combination of these modeling techniques may therefore prove to be the breakthrough required to advance solutions to the current challenges in population genetics of emerging pathogens.

4.5 Materials and Methods

4.5.1 Pangenomic Analysis

The pangenome analysis was run as described in detail before [26]. Briefly, "complete" or "WGS" samples from CC8/ST8 were downloaded from the PATRIC database[24]. Sequences with lengths that were not within 3 standard deviations of the mean length or those with more than 100 contigs were filtered out. A non-redundant list of CDSs from all genomes was created and clustered by protein sequence using CD-HIT with minimum identity and minimum alignment length of 80%[33]. To get the 5' and 3' sequences, non-redundant 300 nucleotide upstream and downstream sequences from the CDS were extracted for each gene.

The CDSs were divided into core, accessory and unique genes based on the frequency of genes as previously described[26]. To calculate the frequency thresholds for each category, $P(x)$, the number of genes with frequency x and its integral $F(x)$, the cumulative frequency less than or equal to x were calculated. The multimodal gene distribution can be estimated by sum of two power laws as:

$$P(x) = c_1 x^{-\alpha_1} + c_2(N + 1 - x)^{-\alpha_2} \qquad x = 1, 2, ..., N$$

where $N$is the total number of genomes, x is the gene frequency and $(c_1, c_2, \alpha_1, \alpha_2, k)$ are parameters fit based on the data. The cumulative distribution $F(x)$ is then the integral of $P(x)$ with additional parameter k:

$$F(x) = k + \frac{c_1}{1 - \alpha_1} x^{1-\alpha_1} - \frac{c_2}{1 - \alpha_2}(N + 1 - x)^{1-\alpha_2}$$

The parameters $(c_1, c_2, \alpha_1, \alpha_2, k)$ were fitted based on the data using non-linear least squares regression from scipy[34]. The frequency threshold of core genomes was defined as greater than $0.9N + 0.1x^*$ and the threshold for unique genome was defined as $0.1x^*$, where x^* represents the inflection point of the fitted cumulative distribution.

4.5.2 Reconstructing the CC8 phylogenetic tree

The phylogenetic tree was reconstructed using the standardized PHaME pipeline on the PATRIC sequences that passed the QC/QA[35]. Using the pipeline, the contigs and sequences were aligned to the reference TCH1516 genome NC_010079 and plasmids NC_012417, NC_010063 [36] and core SNPs were calculated. The core SNPs were then used to estimate the phylogenetic tree using IQ-TREE run with 1000 bootstraps and utilizing the ultrafast bootstrap[37, 38]. The tree was built using the ”TVMe+ASC+G4” model as suggested by the IQ-TREE ModelFinder[39]. Finally, iTOL was used to visualize, annotate and root the tree with the USA100 D592 (NZ_CP035791) from CC5 as the outgroup [40].

4.5.3 Classification of USA300 and non-USA300 strains

The USA300 and non-USA300 strains were classified based on a previously proposed and validated CC8 subtyping scheme[21]. In this scheme, USA300 strains can be identified from the whole genome if they are PVL positive MSSA or MRSA with SCCMec IVa cassette. We detected SCCMec types using SCCMecFinder, and only those genomes where the cassette could be identified by both BLAST and k-mer based methods were marked as positive[22]. PVL was detected using protein BLAST. To find the root of the USA300 strains in the phylogenetic tree, the genomes in the tree were annotated by their PVL and SCCMec status. The tree was then traversed up from reference strain TCH1516 to the CC8 root using ete3, while tracking the total number of genomes, the total number of SCCMec IVa positive genomes and the number of PVL positive genomes in each root[41]. The root of USA300 was placed manually where the number of total genomes kept increasing while the number of PVL and SCCMec positive genomes plateaued. All strains in the clade represented by the USA300 root were classified as USA300 regardless of their SCCMec or PVL status.

4.5.4 DBGWAS and k-mer linkage calculations

DBGWAS was used to enrich mutations unique to USA300 strains. Alleles with frequency less than 0.05 were filtered (-maf 0.05) and all components enriched with q-values less than 0.05 were documented (-SFF q0.05). Genome-wide linkage was estimated by Pearson correlation of the presence/ absence of enriched k-mers and distance was measured based on the k-mer alignment to the reference TCH1516 genome.

To determine the enriched ‘genetic event’ (e.g. SNP, indel, mobile genetic element etc), the graph output from DBGWAS was first loaded onto a networkX model[23]. All nodes in the graph with frequency lower than 0.05 were discarded. MGEs were identified if all significant nodes from DBGWAS had higher frequencies in one strain, e.g. all nodes associated with SCCMec had higher frequencies in USA300 strains. To find SNPs and smaller indel events, the networkx was used to find cycles in the graph, which results from bifurcation and eventual re-collapse of debruijn graphs around mutations. For each cycle, the 'end nodes' representing the start and end of the bifurcation were identified by finding the nodes in the cycle with highest frequency across all samples. As 'end nodes' are present in both case and control samples, they will have higher frequency than other nodes in the cycle which are specific to either case or control. Once the end nodes are identified, the two paths around the bifurcations representing the case and control specific sequences were identified using the shortest path algorithm in networkx. The sequences from nodes of each path were concatenated, changing the sequences to reverse complements and removing overlaps in sequences when required. The concatenated sequences from each path were then compared using BioPython pairwise global alignments to find the SNPs or indels that differentiate the sequences from case and control[42]. If reference sequences are passed, the concatenated sequences are aligned to the reference sequences using BLAST and mutation positions were converted from k-mer positions to positions in the reference genomes. The code used for this analysis can be found in https://github.com/SBRG/dbgwas_network_analysis.

4.5.5 Mapping mutation hotspots with position specific Shannon entropy

For each of the CDS with enriched mutations, the PATRIC local protein family (PLfam) was identified based on the reference TCH1516 genome. All available protein sequences for each CDS PLfam were downloaded and filtered for *S.aureus* sequences. The multilocus sequence type (MLST) of the source genome of each downloaded sequence was mapped using the PATRIC database. The sequences were divided into ST8 and non ST8 and ST239 sequences were filtered. MAFFT was used for multiple alignment and position-specific Shannon entropy was calculated on the aligned file[43]. The entropy was calculated as:

$$H(X) = \sum_{i=1}^{n} P(x_i)log_2P(x_i)$$

where n in the total number of unique amino acids in the position and $P(x_i)$ is the probability of finding the given amino acid.

4.5.6 Calculating strain-specific iModulons with independent component analysis

ICA of RNA sequencing data was performed using the pymodulon package [28]. Using the package, all available RNA sequencing data for NCTC8325 and USA300 strains were downloaded, run through the QC/QA pipeline, manually curated for metadata and aligned to the TCH1516 genome (NC_010079, NC_012417, NC_010063). The combined data was then transformed into log-TPM and normalized to a single reference condition (SRX3760886, SRX3760891). This is in contrast to other ICA models that normalize the data to project specific reference conditions to reduce batch effects. However, normalizing to project specific control conditions also erases the strain specific information as almost all BioProjects contain data from only one isolate (e.g. NCTC8325, TCH1516, LAC etc). ICA was then run as previously described in chapter 3. The activities of the output iModulons were manually parsed to look for iModulons with the largest strain specific differences.

4.5.7 Fur box motif search

Motif search for the Fur box was conducted using FIMO from the online MEME suite[44]. The Bacillus subtilis Fur motif from collecTF was used as a reference[45].

4.6 Acknowledgements

Chapter 4, in part, is currently being prepared for submission for publication: **Poudel, S**, Hyun J, Hefner Y, Nizet V, Palsson B Ø. ”Interpreting roles of mutations in the emergence of S. aureus USA300 strains with genetics and independent component analysis of gene expression.” The dissertation author was the primary author.

4.7 References

1. Young, B. C., Earle, S. G., Soeng, S., Sar, P., Kumar, V., Hor, S., Sar, V., Bousfield, R., Sanderson, N. D., Barker, L., Stoesser, N., Emary, K. R., Parry, C. M., Nickerson, E. K., Turner, P., Bowden, R., Crook, D. W., Wyllie, D. H., Day, N. P., Wilson, D. J. & Moore, C. E. Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. en. *Elife* **8** (Feb. 2019).

2. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V. & Palsson, B. Ø. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

3. Choudhary, K. S., Mih, N., Monk, J., Kavvas, E., Yurkovich, J. T., Sakoulas, G. & Palsson, B. O. The Staphylococcus aureus Two-Component System AgrAC Displays Four Distinct Genomic Arrangements That Delineate Genomic Virulence Factor Signatures. en. *Front. Microbiol.* **9,** 1082 (May 2018).

4. Correction for Copin et al., Sequential evolution of virulence and resistance during clonal spread of community-acquired methicillin-resistant Staphylococcus aureus. en. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 4747 (Mar. 2019).

5. Krishna, A., Holden, M. T. G., Peacock, S. J., Edwards, A. M. & Wigneshweraraj, S. Naturally occurring polymorphisms in the virulence regulator Rsp modulate Staphylococcus aureus survival in blood and antibiotic susceptibility. en. *Microbiology* **164,** 1189–1195 (Sept. 2018).

6. Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., Woodford, N., Smith, E. G., Ismail, N., Llewelyn, M. J., Peto, T. E., Crook, D. W., McVean, G., Walker, A. S. & Wilson, D. J. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. en. *Nat Microbiol* **1,** 16041 (Apr. 2016).

7. Steinig, E. J., Duchene, S., Robinson, D. A., Monecke, S., Yokoyama, M., Laabei, M., Slickers, P., Andersson, P., Williamson, D., Kearns, A., Goering, R. V., Dickson, E., Ehricht, R., Ip, M., O'Sullivan, M. V. N., Coombs, G. W., Petersen, A., Brennan, G., Shore, A. C., Coleman, D. C., Pantosti, A., de Lencastre, H., Westh, H., Kobayashi, N., Heffernan, H., Strommenger, B., Layer, F., Weber, S., Aamot, H. V., Skakni, L., Peacock, S. J., Sarovich, D., Harris, S., Parkhill, J., Massey, R. C., Holden, M. T. G., Bentley, S. D. & Tong, S. Y. C. Evolution and Global Transmission of a Multidrug-Resistant, Community-Associated Methicillin-Resistant Staphylococcus aureus Lineage from the Indian Subcontinent. en. *MBio* **10** (Nov. 2019).

8. Challagundla, L., Reyes, J., Rafiqullah, I., Sordelli, D. O., Echaniz-Aviles, G., Velazquez-Meza, M. E., Castillo-Ramırez, S., Fittipaldi, N., Feldgarden, M., Chapman, S. B., Calderwood, M. S., Carvajal, L. P., Rincon, S., Hanson, B., Planet, P. J., Arias, C. A., Diaz, L. & Robinson, D. A. Phylogenomic Classification and the Evolution of Clonal Complex 5 Methicillin-Resistant Staphylococcus aureus in the Western Hemisphere. en. *Front. Microbiol.* **9,** 1901 (Aug. 2018).

9. Bal, A. M., Coombs, G. W., Holden, M. T. G., Lindsay, J. A., Nimmo, G. R., Tattevin, P. & Skov, R. L. Genomic insights into the emergence and spread of international clones of healthcare-, community-and livestock-associated meticillin-resistant Staphylococcus aureus: blurring of the traditional definitions. *Journal of Global Antimicrobial Resistance* **6,** 95–101 (2016).

10. Uhlemann, A.-C., Dordel, J., Knox, J. R., Raven, K. E., Parkhill, J., Holden, M. T. G., Peacock, S. J. & Lowy, F. D. Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6738–6743 (May 2014).

11. Challagundla, L., Luo, X., Tickler, I. A., Didelot, X., Coleman, D. C., Shore, A. C., Coombs, G. W., Sordelli, D. O., Brown, E. L., Skov, R., Larsen, A. R., Reyes, J., Robledo, I. E., Vazquez, G. J., Rivera, R., Fey, P. D., Stevenson, K., Wang, S.-H., Kreiswirth, B. N., Mediavilla, J. R., Arias, C. A., Planet, P. J., Nolan, R. L., Tenover, F. C., Goering, R. V. & Robinson, D. A. Range Expansion and the Origin of USA300 North American Epidemic Methicillin-Resistant Staphylococcus aureus. en. *MBio* **9** (Jan. 2018).

12. Everitt, R. G., Didelot, X., Batty, E. M., Miller, R. R., Knox, K., Young, B. C., Bowden, R., Auton, A., Votintseva, A., Larner-Svensson, H., Charlesworth, J., Golubchik, T., Ip, C. L. C., Godwin, H., Fung, R., Peto, T. E. A., Walker, A. S., Crook, D. W. & Wilson, D. J. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. en. *Nat. Commun.* **5,** 3956 (May 2014).

13. Thurlow, L. R., Joshi, G. S. & Richardson, A. R. Virulence strategies of the dominant USA300 lineage of community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA). en. *FEMS Immunol. Med. Microbiol.* **65,** 5–22 (June 2012).

14. Nishizaki, S. S., Ng, N., Dong, S., Porter, R. S., Morterud, C., Williams, C., Asman, C., Switzenberg, J. A. & Boyle, A. P. Predicting the effects of SNPs on transcription factor binding affinity. en. *Bioinformatics* **36,** 364–372 (Jan. 2020).

15. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. en. *PLoS One* **7,** e46688 (Oct. 2012).

16. Kavvas, E. S., Yang, L., Monk, J. M., Heckmann, D. & Palsson, B. O. A biochemically-interpretable machine learning classifier for microbial GWAS. en. *Nat. Commun.* **11,** 2580 (May 2020).

17. Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V. & Jacob, L. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. en. *PLoS Genet.* **14,** e1007758 (Nov. 2018).

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

19. Jones, M. B., Montgomery, C. P., Boyle-Vavra, S., Shatzkes, K., Maybank, R., Frank, B. C., Peterson, S. N. & Daum, R. S. Genomic and transcriptomic differences in community acquired methicillin resistant Staphylococcus aureus USA300 and USA400 strains. en. *BMC Genomics* **15,** 1145 (Dec. 2014).

20. Iqbal, Z., Seleem, M. N., Hussain, H. I., Huang, L., Hao, H. & Yuan, Z. Comparative virulence studies and transcriptome analysis of Staphylococcus aureus strains isolated from animals. en. *Sci. Rep.* **6,** 35442 (Oct. 2016).

21. Bowers, J. R., Driebe, E. M., Albrecht, V., McDougal, L. K., Granade, M., Roe, C. C., Lemmer, D., Rasheed, J. K., Engelthaler, D. M., Keim, P. & Limbago, B. M. Improved Subtyping of Staphylococcus aureus Clonal Complex 8 Strains Based on Whole-Genome Phylogenetic Analysis. en. *mSphere* **3** (May 2018).

22. Kaya, H., Hasman, H., Larsen, J., Stegger, M., Johannesen, T. B., Allesøe, R. L., Lemvigh, C. K., Aarestrup, F. M., Lund, O. & Larsen, A. R. SCCmecFinder, a Web-Based Tool for Typing of Staphylococcal Cassette Chromosome mec in Staphylococcus aureus Using Whole-Genome Sequence Data. en. *mSphere* **3** (Jan. 2018).

23. Hagberg, A., Swart, P. & S Chult, D. *Exploring network structure, dynamics, and function using networkx* en. Tech. rep. LA-UR-08-05495; LA-UR-08-5495 (Los Alamos National Lab. (LANL), Los Alamos, NM (United States), Jan. 2008).

24. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E.,

Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic Acids Res.* **42,** D581–91 (Jan. 2014).

25. Robinson, D. A. & Enright, M. C. Evolution of Staphylococcus aureus by large chromosomal replacements. en. *J. Bacteriol.* **186,** 1060–1064 (Feb. 2004).

26. Hyun, J. C., Monk, J. M. & Palsson, B. O. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. en. *BMC Genomics* **23,** 7 (Jan. 2022).

27. Boyle-Vavra, S., Li, X., Alam, M. T., Read, T. D., Sieth, J., Cywes-Bentley, C., Dobbins, G., David, M. Z., Kumar, N., Eells, S. J., Miller, L. G., Boxrud, D. J., Chambers, H. F., Lynfield, R., Lee, J. C. & Daum, R. S. USA300 and USA500 clonal lineages of Staphylococcus aureus do not produce a capsular polysaccharide due to conserved mutations in the cap5 locus. en. *MBio* **6** (Apr. 2015).

28. Sastry, A. V., Poudel, S., Rychel, K., Yoo, R., Lamoureux, C. R., Chauhan, S., Haiman, Z. B., Al Bulushi, T., Seif, Y. & Palsson, B. O. *Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks* en. July 2021.

29. Yuan, Y., Seif, Y., Rychel, K., Yoo, R., Chauhan, S., Poudel, S., Al-bulushi, T., Palsson, B. O. & Sastry, A. *Pan-genomic analysis of transcriptional modules across Salmonella Typhimurium reveals the regulatory landscape of different strains* en. Jan. 2022.

30. Prados, J., Linder, P. & Redder, P. TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens. en. *BMC Genomics* **17,** 849 (Nov. 2016).

31. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. en. *Science* **373,** 871–876 (Aug. 2021).

32. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žıdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. en. *Nature* **596,** 583–589 (Aug. 2021).

33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. en. *Bioinformatics* **28,** 3150–3152 (Dec. 2012).

34. Eric Jones, Travis Oliphant, Pearu Peterson and others. *SciPy: Open Source Scientific Tools for Python* 2001.

35. Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C.-C. & Chain, P. S. G. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. en. *Sci. Rep.* **10,** 1723 (Feb. 2020).

36. Highlander, S. K., Hultén, K. G., Qin, X., Jiang, H., Yerrapragada, S., Mason Jr, E. O., Shang, Y., Williams, T. M., Fortunov, R. M., Liu, Y., Igboeli, O., Petrosino, J., Tirumalai, M., Uzman, A., Fox, G. E., Cardenas, A. M., Muzny, D. M., Hemphill, L., Ding, Y., Dugan, S., Blyth, P. R., Buhay, C. J., Dinh, H. H., Hawes, A. C., Holder, M., Kovar, C. L., Lee, S. L., Liu, W., Nazareth, L. V., Wang, Q., Zhou, J., Kaplan, S. L. & Weinstock, G. M. Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. en. *BMC Microbiol.* **7,** 99 (Nov. 2007).

37. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. & Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. en. *Mol. Biol. Evol.* **37,** 1530–1534 (May 2020).

38. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. en. *Mol. Biol. Evol.* **35,** 518–522 (Feb. 2018).

39. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. en. *Nat. Methods* **14,** 587–589 (June 2017).

40. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. en. *Nucleic Acids Res.* **49,** W293–W296 (July 2021).

41. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. en. *Mol. Biol. Evol.* **33,** 1635–1638 (June 2016).

42. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. en. *Bioinformatics* **25,** 1422–1423 (June 2009).

43. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. en. *Nucleic Acids Res.* **30,** 3059–3066 (July 2002).

44. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME SUITE: tools for motif discovery and searching. en. *Nucleic Acids Res.* **37,** W202–8 (July 2009).

45. Kılıç, S., White, E. R., Sagitova, D. M., Cornish, J. P. & Erill, I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. en. *Nucleic Acids Res.* **42,** D156–D160 (Nov. 2013).

Chapter 5

Conclusions

The accelerating accumulation of sequencing data presents an unprecedented opportunity to harness the power of scale to address the challenges of infectious diseases and antimicrobial resistance. However, new modeling systems and analytical tools are required to take full advantage of this emerging resource. In this dissertation, we presented one such approach. We used ICA to model the TRN of *S. aureus* USA300 strain from the growing number of publicly available RNA sequencing data. Overall, we show that this model of the TRN is scalable, interpretable and functional which makes it ideal to study non-model organisms where data and available resources can be more sparse. The size and the coverage of the model grows with the increasing number of publicly available data, enabling rapid characterization of new regulatory structures and dynamics. The ensuing model generated by ICA can be used to understand the TRN from gene to genome scale and aides in interpreting complex expression profiles and regulatory interactions. Lastly, the model can be used in conjunction with existing modeling approaches, extending the functionality of both methods. Taking advantage of this flexible modeling system, we characterized the structure, dynamics and evolution of *S. aureus* USA300 strains.

The second chapter of this dissertation describes the first ICA model of the TRN that was built for this organism. The model was used to interpret complex expression profiles from infection mimicking conditions, define the regulatory role of poorly understood alternative sigma factor SigS, discover new potential interactions between expression of prophages and virulence and connect metal requirements with carbon metabolism. The model opened the door for using ICA to accelerate the characterization of the structure of TRN in *S. aureus*.

The third chapter explores transcriptional regulation of metabolism. Our work showed that CcpA and CodY, the regulators of central carbon and amino acid metabolism respectively, coordinated their activity. By integrating the TRN structure provided by iModulons with metabolic models, we showed that this coordination also exists at the level of metabolic fluxes. Furthermore, we show that these interactions extend to expression of translation associated genes, thus integrating metabolic signals with protein synthesis. We combined these observations into a single coarse grained model of coordinated protein synthesis regulation captured by CcpA, CodY and Translation iModulons.

The fourth chapter demonstrates how ICA can be used to extend the limitations of genetic approaches when studying emerging endemic strains. Here, we showed how inferring the role of mutations contributing to the success of USA300 strains with traditional gene association studies can be limited by strong genetic linkage and population structure. However, by modeling gene regulation with ICA, we found several genes with enriched mutations that were also differentially regulated in a strain-specific manner. In line with this observation, we found large changes in the regulatory region of *isdH* gene in USA300 strains, including the deletion of Fur transcription factor binding sites. This opens up new possibilities for combining genetic and transcriptomic data to study mutations important for clinical success of new strains.

Current systems biology and bioinformatics stands at the cusp of the next big revolution like the ones brought on by the first draft of human genome or the first description of the structure of DNA. This time, the acceleration in research and rapid expansion of knowledge will be brought on by the emergence of new analysis and modeling methods and fueled by the concurrent exponential growth in new biological data. Already new techniques, often derived from the field of machine learning, have made great strides towards designing new drugs, solving the protein folding problem, processing biological images, and deconvoluting high dimensional nonlinear biological data. This work makes a small contribution towards this exciting new era in the field of systems biology.

Appendix A

Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response

A.1 Supplementary Figures

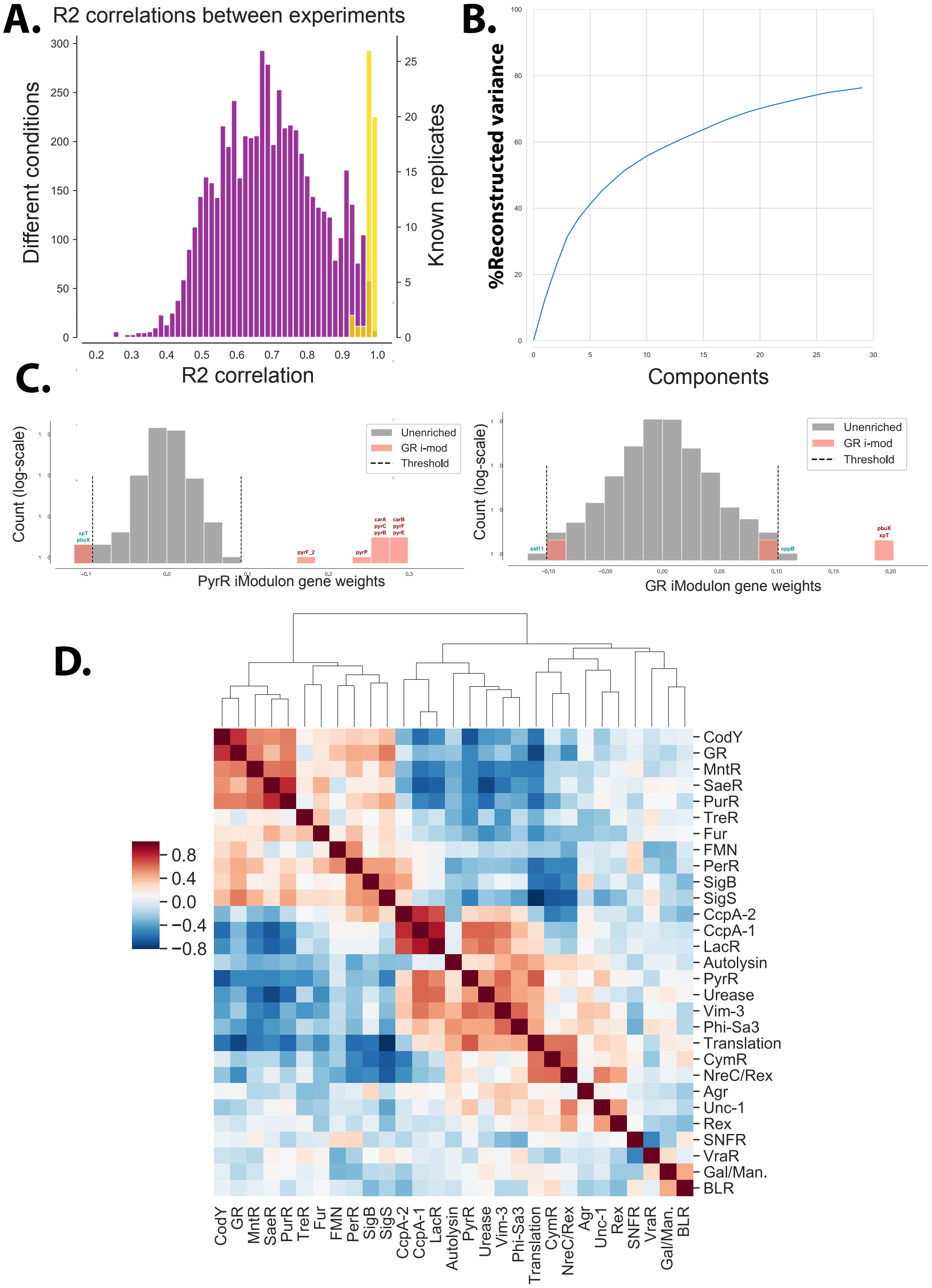


Figure A.1: Caption in the next page.

Mathematical representation of *S.aureus* TRN. (a) RNA sequencing data were collected in duplicates and their reproducibility was verified using Spearman correlation of TPM values. Correlation between replicates (yellow bars) for all samples had r-squared 0.9, with most samples having r-squared 0.95. Correlation between different samples (purple bars) had a wide range of correlation, indicating the presence of diverse expression states. (b) The ICA decomposition captured most of the information in the input RNA sequencing compendium (Dataset S7). 76% of the total variance could be reconstructed from the product of S (Dataset S8) and A (Dataset S9). (c) Histogram of gene coefficient in two example components (containing iModulon for pyrmindine above and GR below). While most genes in a component have weights close to 0, few statistically significant outliers (outside of the vertical dashed lines) with high weightings (red bars) form an independently modulated set of genes (called an iModulon). Genes can have both positive and negative coefficients and can be present in multiple iModulons. The genes *xpt* and *pbuX* have negative coefficient in the PyrR iModulon (top histogram) indicating that these genes are contra-regulated to genes with positive coefficient in the same iModulon(.e.g *carAB*). *Xpt* and *pbuX* are also present in the GR iModulon (bottom histogram), indicating that these two genes are regulated by multiple regulators. The first row of the matrix also contains the threshold used to call iModulons. (d) Though iModulons represent independently regulated set of genes, their activities are coordinated with one another. The coordination is visualized as a heatmap depicting Pearson correlation of iModulon activities across all 108 samples.

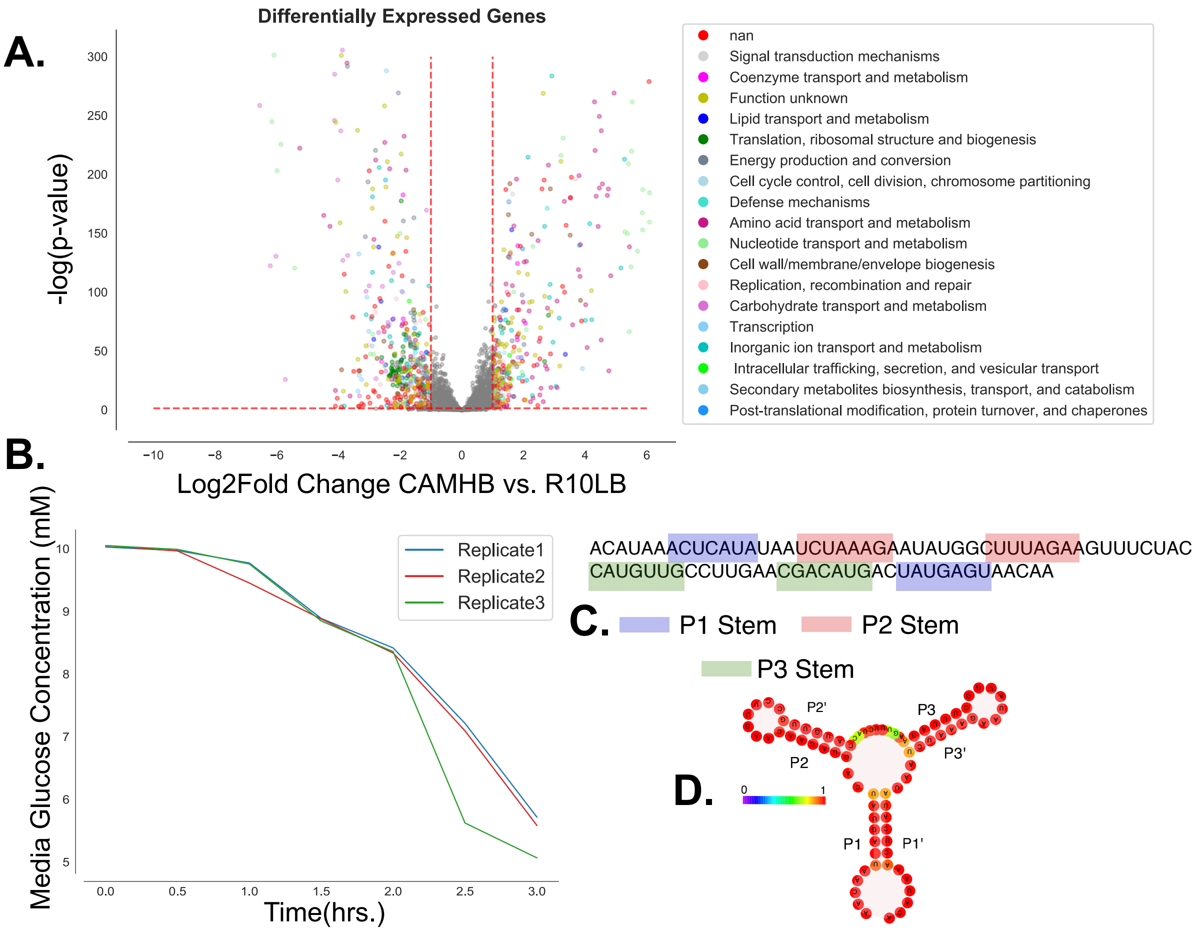


Figure A.2: Differential activation analysis and verification. (a) Volcano plot of differential expression levels of genes between CAMHB and R10LB. 848 genes spanning at least 17 COG categories (as determined by EggNog v4.5 [1]) were significantly differentially expressed. Genes with greater than 2-fold change in expression and with p-value ¡ 0.05 were considered significantly differentially expressed. (b) Glucose uptake was measured in R10LB and CAMHB. *S.aureus* actively took up glucose in R10LB while no glucose was detected in CAMHB. Each line represents a biological replicate in R10LB. (c) Riboswitch in conserved sequence upstream of *xpt* gene was verified using riboswitch finder[2]. (d) The structure of the riboswitch was verified with RNAfold within the ViennaRNA Package 2.0. [3]

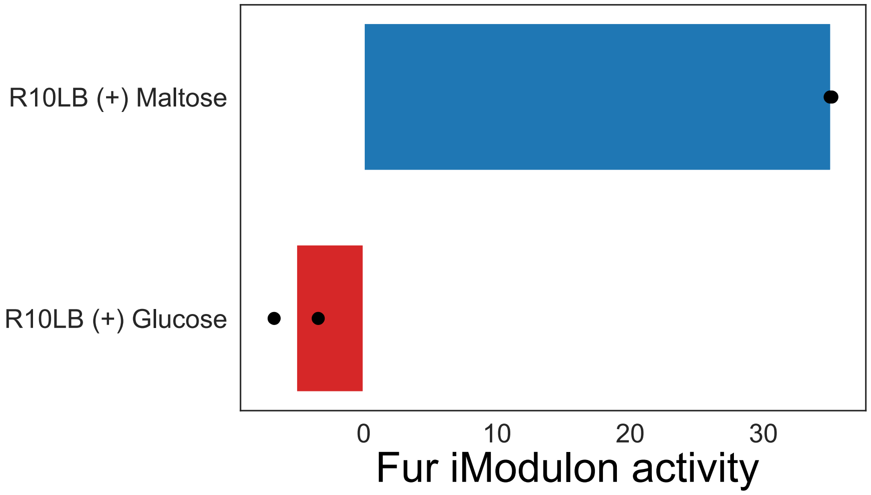


Figure A.3: Fur activity in response to changes in carbon source. (a) The activity of Fur iModulons increased when the carbon source in R10LB was changed from glucose to maltose

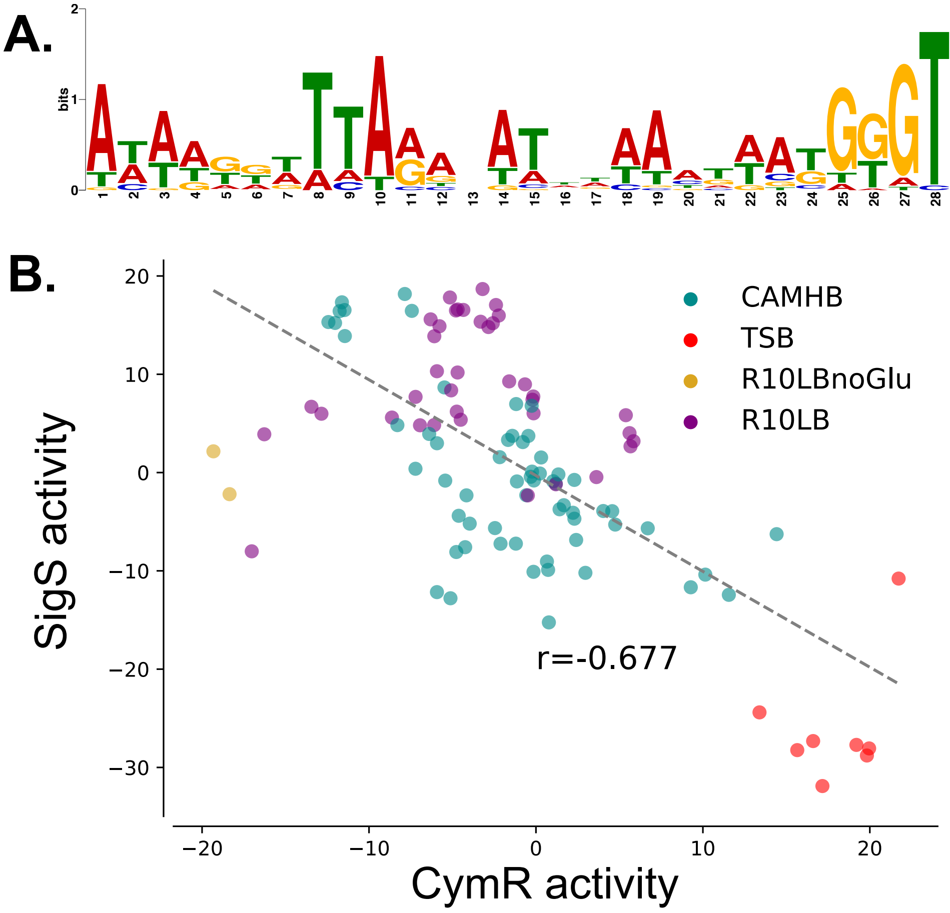


Figure A.4: Sigma Factor iModulons.(a) Regulatory region of SigB iModulon contained a conserved motif that closely matched SigB motif of *B. subtilis*. (b) The activity level of SigS iModulon was negatively correlated with the activity of CymR iModulon (PearsonR=0.677, p-val=8.29e-16).

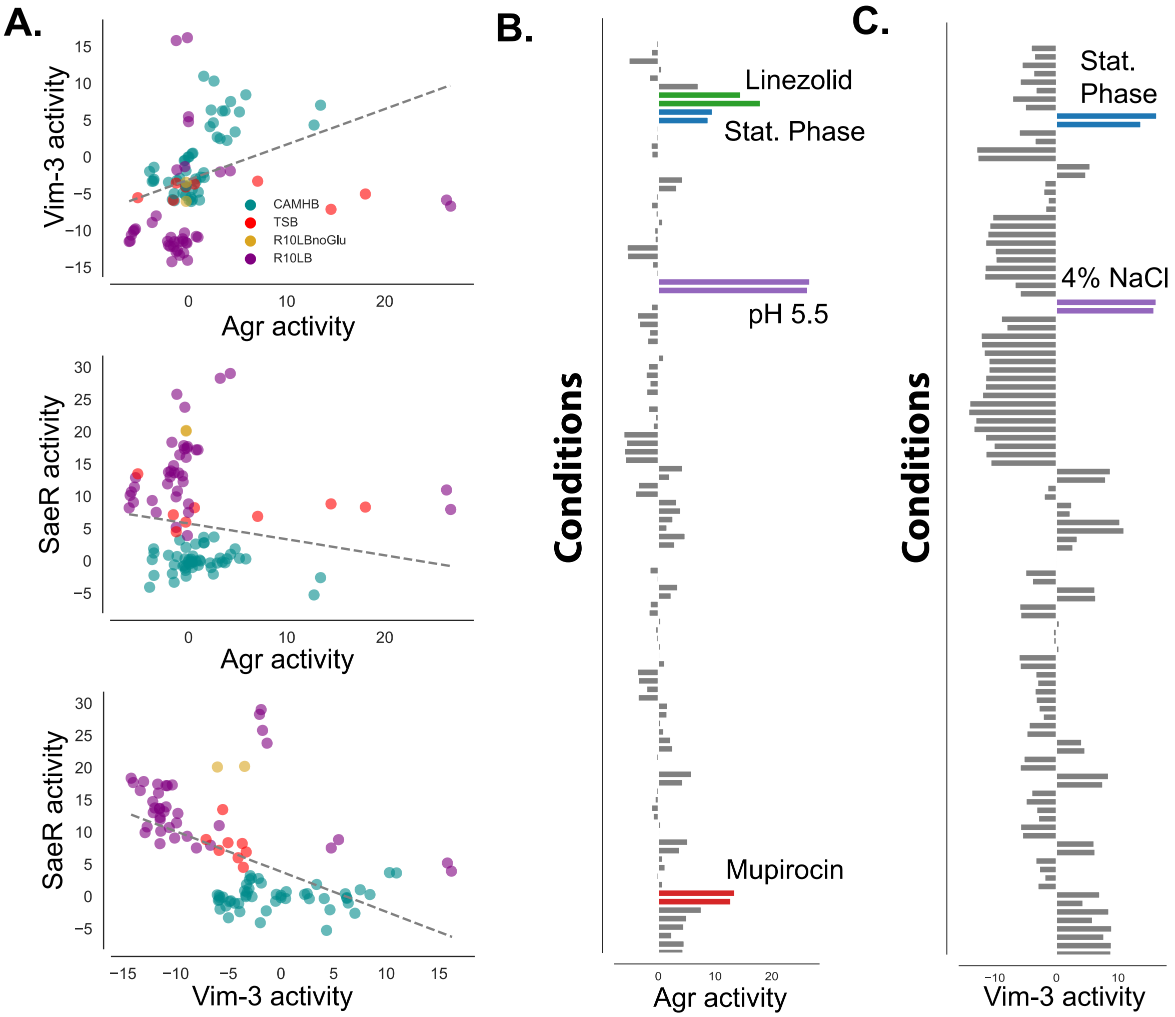


Figure A.5: The virulence iModulons of *S.aureus*. (a) Activity of Agr was poorly correlated with the activities of the other two virulence associated iModulons, SaeR and Agr. However, SaeR and Vim-3 activities were negatively correlated. (b) Agr activity in most samples were close to 0. Its activity could be induced by translation inhibitors, growth to OD600 of 1 (Stat. Phase), and low pH (5.5). (c) The Vim-3 iModulon had the highest activity in stationary phase and when *S.aureus* was challenged with 4% NaCl.

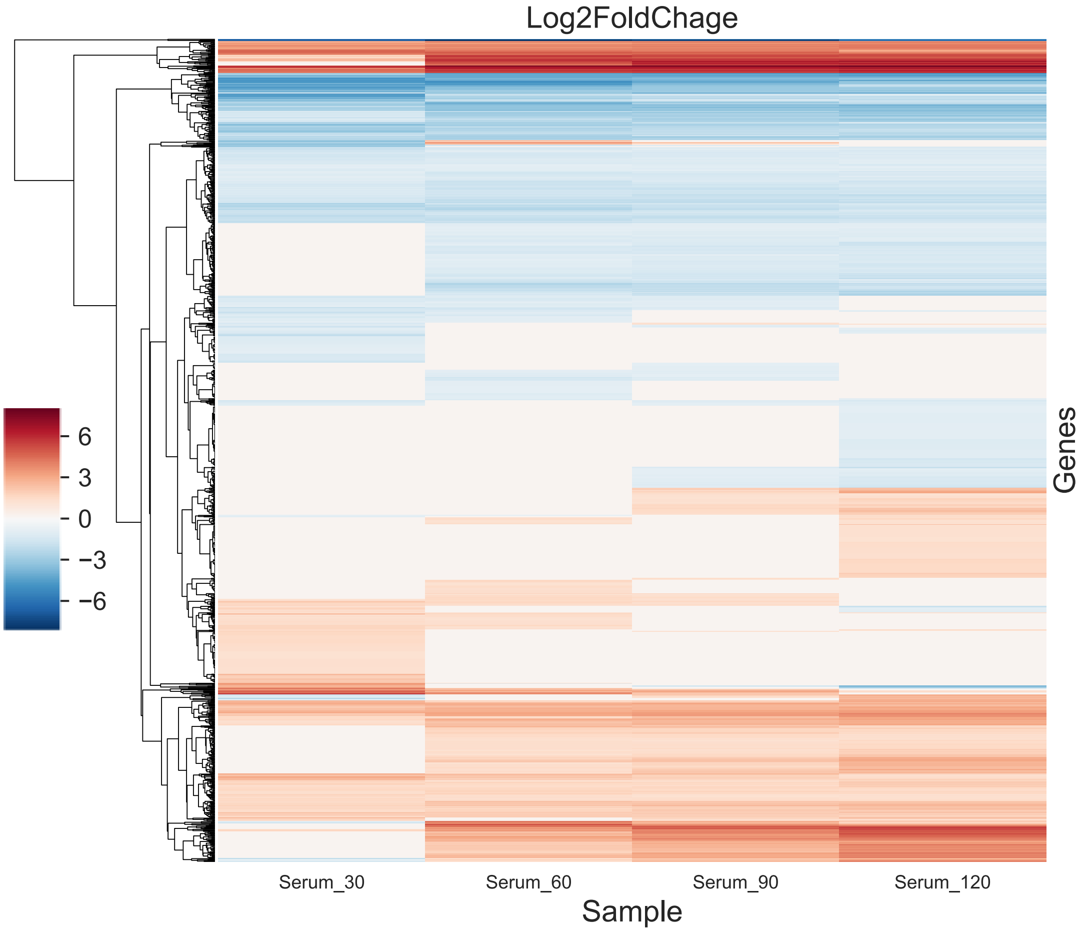


Figure A.6: Differentially expressed genes in Serum. Clustermap of 1177 genes that were differentially expressed in at least one of the serum samples.

A.2 Analysis of complex in vitro and ex vivo expression profiling data with ICA and StaphNet

ICA analysis provides a low dimensional and biologically relevant decomposition of expression profiling data. This decomposition recasts the expression data into activity of independently modulated sets of genes (iModulons), making the data far more interpretable. To demonstrate this, we reanalyzed the ex vivo serum data from Figure 6 using a graph based model of *S.aureus* TRN named StaphNet[4], and traditional differential expression analysis. Our analysis demonstrates that output of ICA analysis is more interpretable than those provided by differential expression analysis or by StaphNet.

StaphNet is a probabilistic functional gene network of USA300 strain FPR3757 built by combining genomic data from multiple sources. Users can use this model to explore their differential expression data using a method called 'Context-centric Search.' Given a set of differentially expressed genes (DEGs) this algorithm finds hub genes (genes with $=$ 20 connections) which have neighbors that are significantly overlapped with input DEGs. This allows the users to understand which genetic hubs the DEGs are centered around. Unlike ICA analysis, StaphNet does not provide any form of activity or expression levels as output and therefore cannot be used to generate time-series data as presented in Figure 6a. On the other hand, while differential expression analysis gave gene expression levels for each of the time-point, each time point had over 100 DEGs which could not be conveyed clearly in a time-series plot. Therefore, we chose to compare the 2 hour time point. Traditional comparison of gene expression after two hour growth in serum revealed that at this time-point there were 848 genes spanning dozen COG categories that were differentially expressed which made it difficult to fully characterize the response of *S.aureus*. Analyzing the top 500 DEGs with StaphNet (the maximum number allowed by the algorithm) yielded at least 100 gene hubs that were enriched in proximity to the DEGs (Dataset S10). These hubs are ranked by StaphNet and the products of the top 5 hub genes were DNA-directed RNA polymerase subunit delta (SAUSA300_RS14555), polysaccharide deacetylase (SAUSA300_RS14530), DUF3816 family protein (SAUSA300_RS14550), L-threonine dehydratase biosynthetic IlvA (SAUSA300_RS11075), dihydroxy-acid dehydratase (SAUSA300_RS11035) (Table1). In contrast, ICA analysis provided clear differential activation of different regulators in the serum (e.g. SaeR, AgrA, CodY, Fur).The analysis also outputs the activity of each of these regulator associated iModulons, which allows us to follow their dynamics through the time-course. For example, while both Fur and CodY activity are very high in serum at 2 hour time-point, Fur activity jumps immediately when introduced to serum while the CodY activity steadily increases over time to match Fur by 2 hours (Figure 6a). Indeed these dynamics not cannot be readily inferred from the expression levels of 1177 genes that were differentially expressed in at least one of the serum time-points (Figure S6).

A.2.1 Methods

For methods used for ICA analysis of serum data, please see the main text. The differentially expressed genes and their expression level in serum was used as reported by the original paper[5]. The top 500 differentially expressed genes in Serum at 2 hour time-point was submitted to the online implementation of the StaphNet Context-centric Search algorithm (https://www.inetbio.org/staphnet/Network_regulon_form.php). The products of the top 5 hub genes were determined using Aureowiki (Table 1)[6].

A.3 Supplementary Tables

Table A.1: StaphNet context-specific search top hits

Rank	USA300_FPR3757 locus tag	*S. aureus* GO terms	p-value
1	SAUSA300_RS14555		6.65E-28
2	SAUSA300_RS14530	GO:0005975-carbohydrate metabolic process	2.02E-27
3	SAUSA300_RS14550		4.20E-27
4	SAUSA300_RS11075	GO:0009097-isoleucine biosynthetic process,GO:0006566-threonine metabolic process	2.13E-25
5	SAUSA300_RS11035	GO:0009097-isoleucine biosynthetic process, GO:0009099-valine biosynthetic process	3.31E-23

Top 5 hubs (degree = 20) enriched from the top 500 differentially expressed genes in serum at 2 hour time-point. The hubs were determined using the 'Context-centric Search' method from StaphNet.

A.4 References

1. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. en. *Nucleic Acids Res.* **44,** D286–93 (Jan. 2016).

2. Bengert, P. & Dandekar, T. Riboswitch finder–a tool for identification of riboswitch RNAs. en. *Nucleic Acids Res.* **32,** W154–9 (July 2004).

3. Lorenz, R., Bernhart, S. H., Siederdissen, C. H. z., Tafer, H., Flamm, C., Stadler, P. F. & Hofacker, I. L. *ViennaRNA Package 2.0* 2011.

4. Kim, C. Y., Lee, M., Lee, K., Yoon, S. S. & Lee, I. Network-based genetic investigation of virulence-associated phenotypes in methicillin-resistant Staphylococcus aureus. en. *Sci. Rep.* **8,** 10796 (July 2018).

5. Malachowa, N., Whitney, A. R., Kobayashi, S. D., Sturdevant, D. E., Kennedy, A. D., Braughton, K. R., Shabb, D. W., Diep, B. A., Chambers, H. F., Otto, M. & DeLeo, F. R. Global changes in Staphylococcus aureus gene expression in human blood. en. *PLoS One* **6,** e18617 (Apr. 2011).

6. Fuchs, S., Mehlan, H., Bernhardt, J., Hennig, A., Michalik, S., Surmann, K., Pané-Farré, J., Giese, A., Weiss, S., Backert, L., Herbig, A., Nieselt, K., Hecker, M., Völker, U. & Mäder, U. AureoWiki The repository of the Staphylococcus aureus research and annotation community. en. *Int. J. Med. Microbiol.* **308,** 558–568 (Aug. 2018).

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

8. Sastry, A. V., Poudel, S., Rychel, K., Yoo, R., Lamoureux, C. R., Chauhan, S., Haiman, Z. B., Al Bulushi, T., Seif, Y. & Palsson, B. O. *Mining all publicly available expression data to compute dynamic microbial transcriptional regulatory networks* en. July 2021.

9. Krueger, F. Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* **516** (2015).

10. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data (2010).

11. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (Mar. 2012).

12. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. en. *Bioinformatics* **31,** 166–169 (Jan. 2015).

13. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. en. *Bioinformatics* **32,** 3047–3048 (Oct. 2016).

14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

15. Koldovsky, Z., Tichavsky, P. & Oja, E. Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the CramÉr-Rao Lower Bound. *IEEE Trans. Neural Netw.* **17,** 1265–1277 (Sept. 2006).

16. McConn, J. L., Lamoureux, C. R., Poudel, S., Palsson, B. O. & Sastry, A. V. Optimal dimensionality selection for independent component analysis of transcriptomic data. en. *BMC Bioinformatics* **22,** 584 (Dec. 2021).

17. Ravcheev, D. A., Best, A. A., Tintle, N., Dejongh, M., Osterman, A. L., Novichkov, P. S. & Rodionov, D. A. Inference of the transcriptional regulatory network in Staphylococcus aureus by integration of experimental and genomics-based evidence. en. *J. Bacteriol.* **193,** 3228–3240 (July 2011).

Appendix B

Coupling of CcpA and CodY activities coordinates carbon and nitrogen metabolism associated gene expression in *S. aureus* USA300 strains

B.1 Supplementary Figures

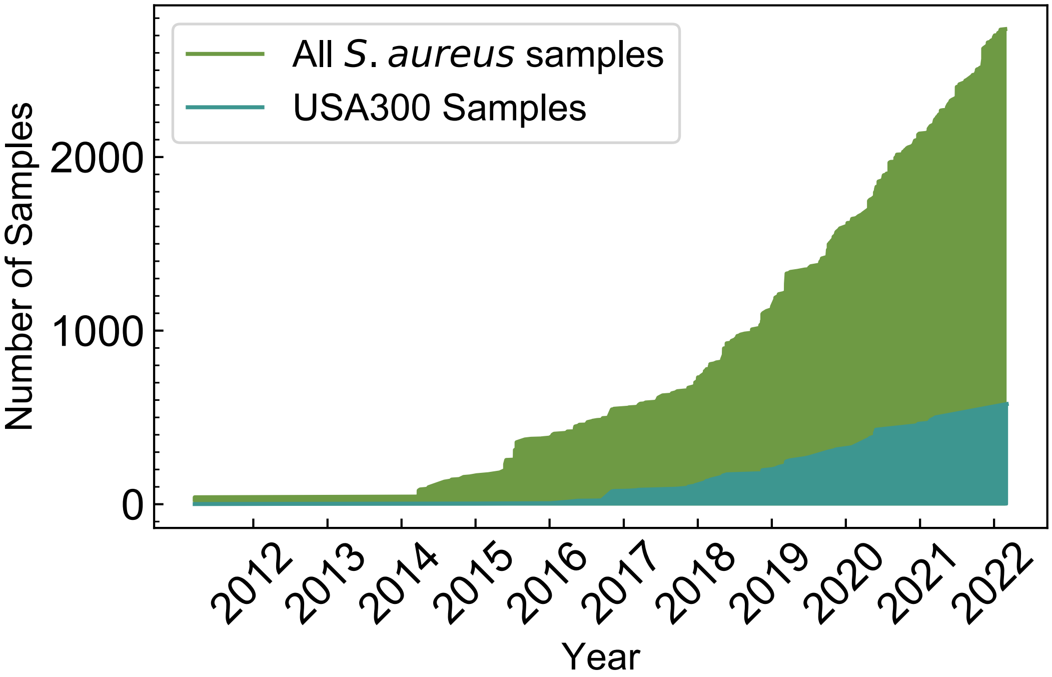


Figure B.1: Growth of available *S. aureus* and USA300 strain specific samples available in SRA.

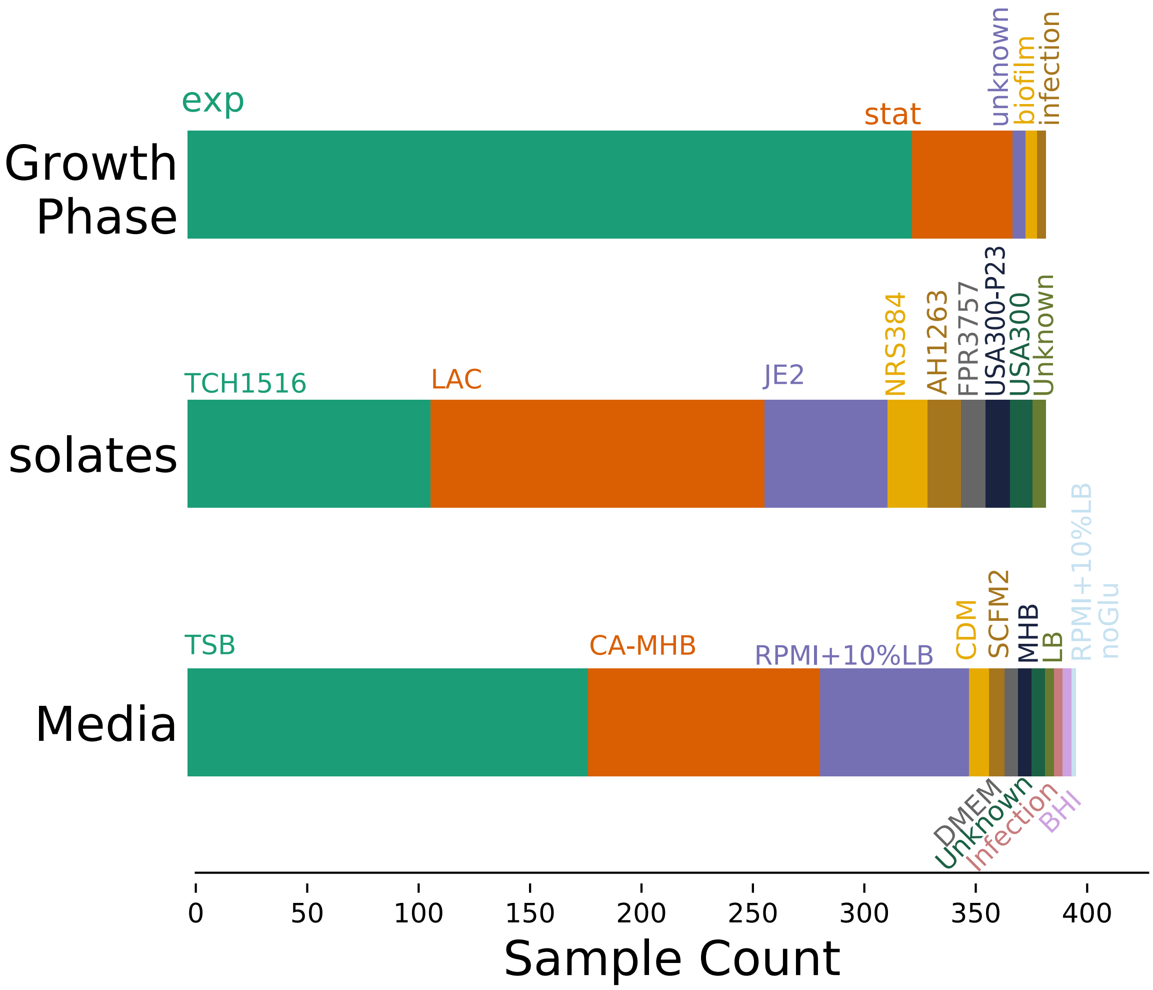


Figure B.2: Staphylococcus aureus RNA sequencing sample distribution based on SRA metadata and manual curation from linked publications, if available.

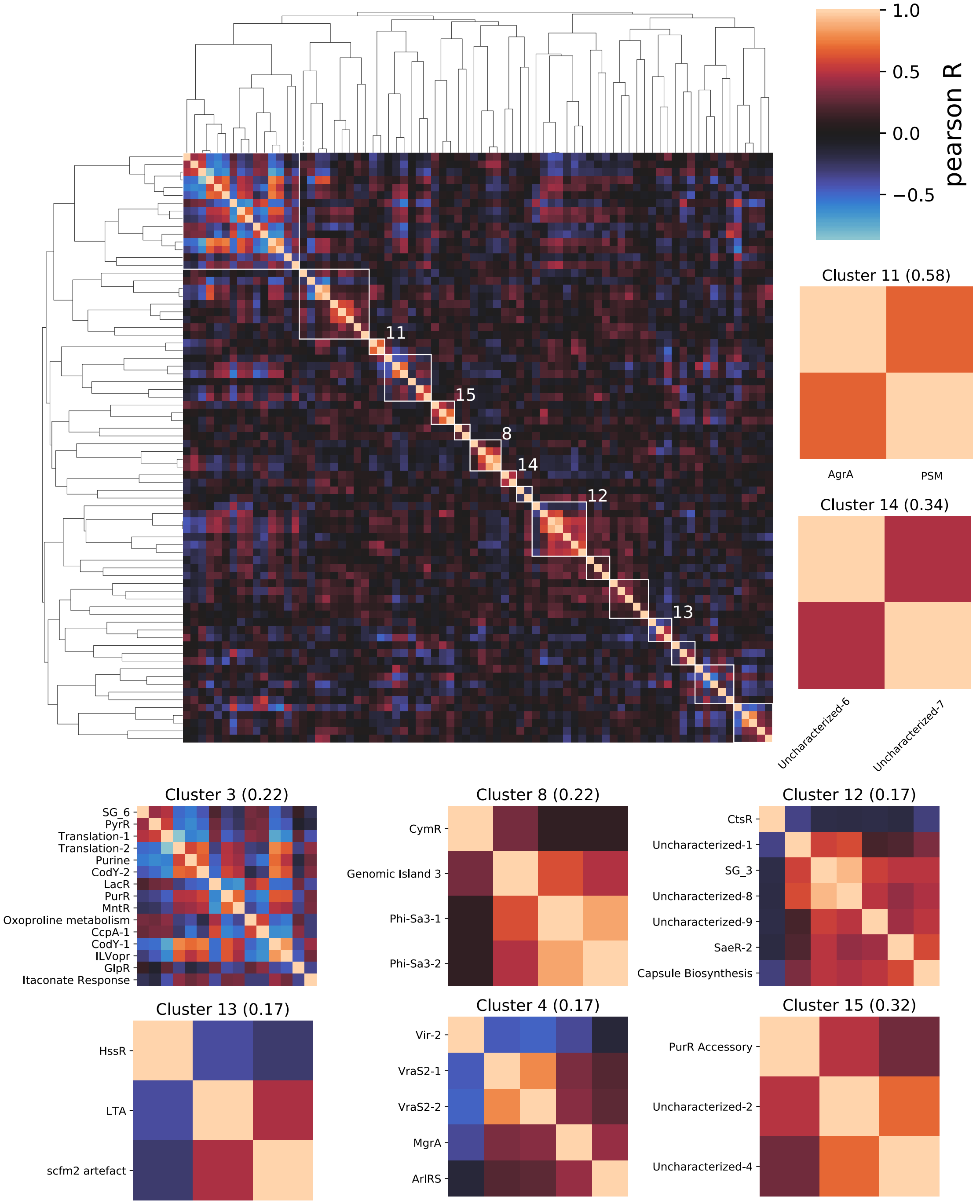


Figure B.3: The iModulon activities formed distinct clusters indicating coordinated gene regulation. The highlighted individual clusters are presented in the smaller clustermaps.

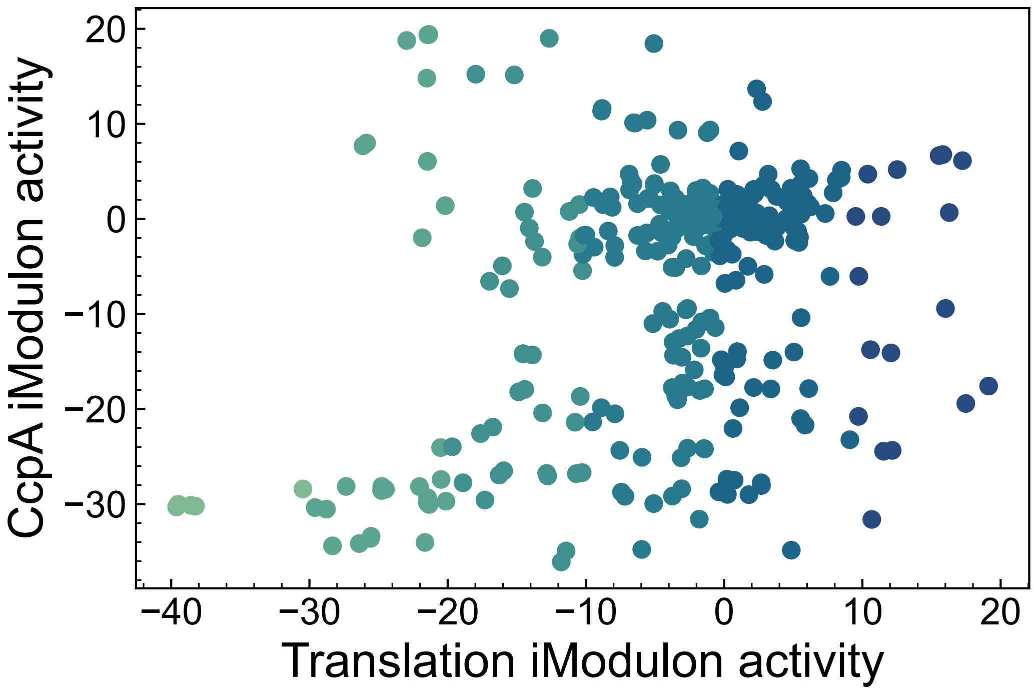


Figure B.4: The iModulon activities of CcpA and Translation activities showed little correlation.

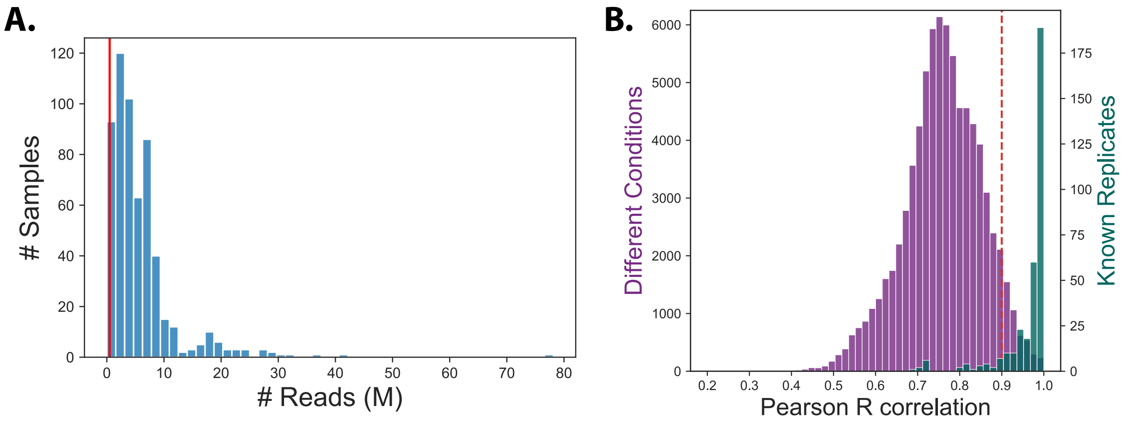


Figure B.5: RNA sequencing quality control metrics. (a) Samples with less that 500,000 reads were filtered out as were samples with Pearson correlation of less than 0.9 (b).

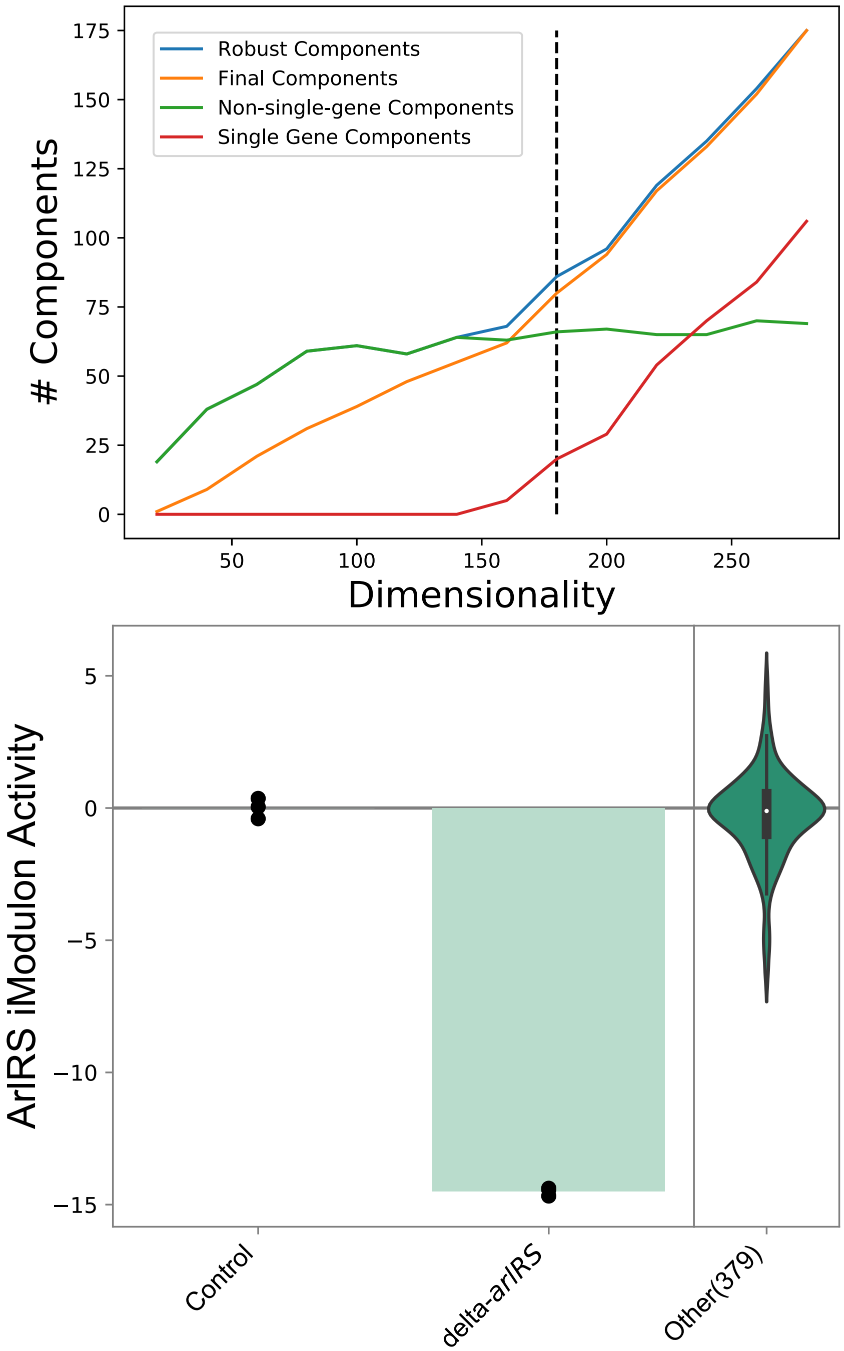


Figure B.6: iModulon post-processing.(a) ICA models with different dimensionality were created with OptICA. Model using 180 dimensions which had the maximum number of robust components while still keeping single gene components to a minimum was chosen as the final model. (b) Some iModulons were labeled based on data from gene knockout experiments. Here, iModulon with lowest activity in delta arlRS strain was labeled as ArlRS iModulon.

B.2 Creating the iModulon model for TRN of *S. aureus* USA300 data

This section provides detailed information on the steps taken to build the iModulon model of *S. aureus* USA300 strains. The methods described in this section were developed previously in *E. coli*, which contains further details and explanations[7, 8]. We started the process by downloading metadata for all the available *S. aureus* RNA sequencing data in the Short Reads Archive (SRA). We then curated the metadata manually to separate samples from a USA300 lineage (e.g. TCH1516, LAC, FPR3757 etc) from all other strains. Fastq files associated with these samples were downloaded from SRA, trimmed with TrimGalore and aligned to the TCH1516 reference genome, including the two plasmids (NC_010079, NC_012417, NC_010063) using Bowtie2[9–11]. The gene read counts were then determined using HTSeqCount with intersection-strict criteria. The counts were then normalized and transformed to create log2TPM[12].

The quality of the alignment was checked using fastqc and any data failing 'per base sequence quality,' 'per sequence quality score,' 'per base n content', or 'adapter content' were dropped. We also dropped samples with less than 500,000 reads aligned to one of the known CDS in TCH1516 (Figure B.5a). These QC stats were organized into a single metadata using MultiQC[13]. Samples that had poor correlation with other samples or clustered with samples from different projects were also excluded. For all samples that passed these QC steps, we searched through online records including SRA, BioSample or linked publications to gather additional information such as base media, growth conditions, mutations etc. We discarded samples that had little or no metadata.

All samples were assigned to a specific project based on the source of the data i.e. all samples from the same BioProject or publication were assigned to the same project. The samples in all projects were checked for reproducibility by checking the Pearson correlation between log2TPM of replicates. Samples with no replicates or those with r-value less than 0.9 were excluded at this step (Figure B.5b). For each project, we identified one reference or control condition. This condition was used to center the data, by subtracting the log2TPM value of the reference from all other conditions in the same project. This reduces iModulons associated with inter-project batch effects. It also sets the log2TPM and all iModulon activities in reference conditions to 0 which allows us to easily interpret activity of iModulons in other samples as fold change from the control. The final 385 samples that passed these QC/QA steps were used to calculate iModulons.

We applied FastICA implemented in the scikit-learn package to calculate the M and A matrix from the logTPM data[14, 15]. FastICA was applied 100 times with random seed and identical components from each run (which may contain slightly different values) were identified after clustering with DBSCAN. Only components that appeared in each run were kept for further analysis. Unlike PCA, the number of components that ICA calculates is not fixed and is a required input in FastICA. Decomposing the transcriptome into too few components can lead to signals from multiple regulators being combined into one iModulon. On the other hand, too many components leads to over decomposition that results in iModulons with a single gene or iModulons with near 0 activity in all samples. To determine the ideal number of components, we used our previously developed OptICA method[16]. OptICA runs ICA with 10 to 340 components with 10 component increments as inputs. For each model, with different component number input, we checked the number of robust and single gene iModulons. For the final model, we chose iModulons calculated with 170 components as it maximized robust components while minimizing the number of single gene iModulons(Figure B.6a).

Once the model with optimal dimensionality was identified, we annotated the iModulons. iModulons were first annotated by comparing the enriched genes in each component to other predicted regulons from regPRECISE and other literature sources(see ‘TRN’ object in the model)[17]. iModulons with significant overlap with predicted regulons; significant overlap was defined as hypergeometric test p-value ¡0.05, precision ¿= 0.5 and coverage ¿= 0.2. However, we also manually curated iModulons as not all regulators have predicted regulons and ICA can predict iModulons that are associated with other biological features (e.g. plasmids, prophages, gene deletions etc). ‘Functional’ iModulons were named after the functions of enriched genes in them e.g. Translation, Autolysins and Beta Lactam Resistance. In cases where data from regulator deletion mutants were available, iModulons were named if they showed the highest change in activity in the mutants (Figure B.6b).

B.3 Supplementary Tables

Table B.1: Metabolites at the intersection of CcpA and CodY regulated metabolism

Metabolite	CodY Reactions	CcpA Reactions
Tetrahydrofolate	Methionine synthase	Glycine cleavage complex

Metabolite	CodY Reactions	CcpA Reactions

2-Oxoglutarate	Phenylalanine transaminase;
	Glutamate synthase;
	Phosphoserine transaminase;
	Tyrosine transaminase;
	N-acetyl-LL-diaminopimelate aminotransferase;
	4-aminobutyrate transaminase;
	Histidinol-phosphate transaminase reversible;
	3-Aminopropanoate 2-oxoglutarate aminotransferase
		Glutamate dehydrogenase;
		2-Oxoglutarate dehydrogenase;
		Oxalosuccinate carboxy-lyase;
		Ornithine transaminase;
		Succinyldiaminopimelate transaminase

L-Phenylalanine	Phenylalanine transaminase	L phenylalanine transporter
Acetaldehyde	Ethanol NAD oxidoreductase	Deoxyribose-phosphate aldolase
Sl2a6[1]	Tetrahydrodipicolinate succinylase	Succinyldiaminopimelate transaminase
Glycerol	Glycerol Dehydrogenase	Glycerophosphodiester phosphodiesterase; Glycerol kinase; Glycerol symporter
L-Threonine	Threonine synthase; L-threonine deaminase	L-threonine dehydrogenase
L-Histidine	L-Histidinal NAD oxidoreductase	Histidase

Metabolite	CodY Reactions	CcpA Reactions
Succinyl-CoA	Tetrahydrodipicolinate succinylase	Succinyl-CoA synthetase; 2-Oxoglutarate dehydrogenase
5,10-mTHF[1]	5,10-methylenetetrahydrofolatereductase	Glycine cleavage complex
L-Alanine	Alanine-Sodium symporter; L-Alanine-proton symporter	L-alanine dehydrogenase
Gly-3-p[1]	Tryptophan synthase (indoleglycerol phosphate)	Deoxyribose-phosphate aldolase reversible; Glyceraldehyde-3-phosphate dehydrogenase (NADP)
Glycine	Glycine-proton symporter	Glycine-cleavage complex; Glycine C-acetyltransferase
L-Aspartate	Aspartate kinase; L-Aspartate 2-oxoglutarate aminotransferase	Aspartate-Sodium symporter

Metabolite	CodY Reactions	CcpA Reactions
Pyruvate	Anthranilate synthase; Anthranilate synthase, ammonia; Dihydrodipicolinate synthase; Cystathionine b-lyase	Maltotriose transport via PTS; L-alanine dehydrogenase; Pyruvate kinase; Pyruvate formate lyase; L-ascorbate transport via PEP:Pyr PTS ; N-Acetylneuraminate lyase; Dihydroxyacetone phosphotransferase; Trehalose transport via PEP:Pyr PTS
Oxaloacetate	L-Aspartate 2-oxoglutarate aminotransferase	Citrate synthase; Phosphoenolpyruvate carboxykinase
L-Tryptophan	Tryptophan synthase (indole)	L-tryptophan-proton symporter
L-Arginine	L-Arginine transporter	Arginase
Acetyl-CoA	Homoserine O-trans-acetylase; Acetyl-CoA L-2, 3, 4, 5-tetrahydrodipicolinate N2-acetyltransferase	Pyruvate formate lyase; Citrate synthase; Glycine C-acetyltransferase
L-Tyrosine	Tyrosine transaminase	L-tyrosine-proton symporter

Metabolite	CodY Reactions	CcpA Reactions
D-Glyceraldehyde	Glycerol Dehydrogenase	Glyceraldehyde facilitated diffusion

L-Glutamate	Phenylalanine transaminase;
	Glutamate synthase;
	Glutathione hydralase;
	Phosphoserine transaminase;
	Anthranilate synthase;
	Tyrosine transaminase;
	N-acetyl-LL-diaminopimelate aminotransferase;
	4-aminobutyrate transaminase;
	Histidinol-phosphate transaminase reversible;
	3-Aminopropanoate 2-oxoglutarate aminotransferase;
	4-amino-4-deoxychorismate synthase;
	Imidazole-glycerol-3-phosphate synthase
		Glutamate dehydrogenase;
		Glutamate-Sodium symporter;
		1-pyrroline-5-carboxylate dehydrogenase;
		Ornithine transaminase reversible;
		Succinyldiaminopimelate transaminase

[1] Abbreviations: **Gly-3-P**:Glyceraldehyde 3-phosphate; **5,10-mTHF**:5,10-methylTetrahydrofolate; **Sl2a6**: N-Succinyl-2-L-amino-6-oxoheptanedioate

Appendix C

Interpreting roles of mutations in the emergence of *S. aureus* USA300 strains with genetics and independent component analysis of gene expression

C.1 Supplementary Figures

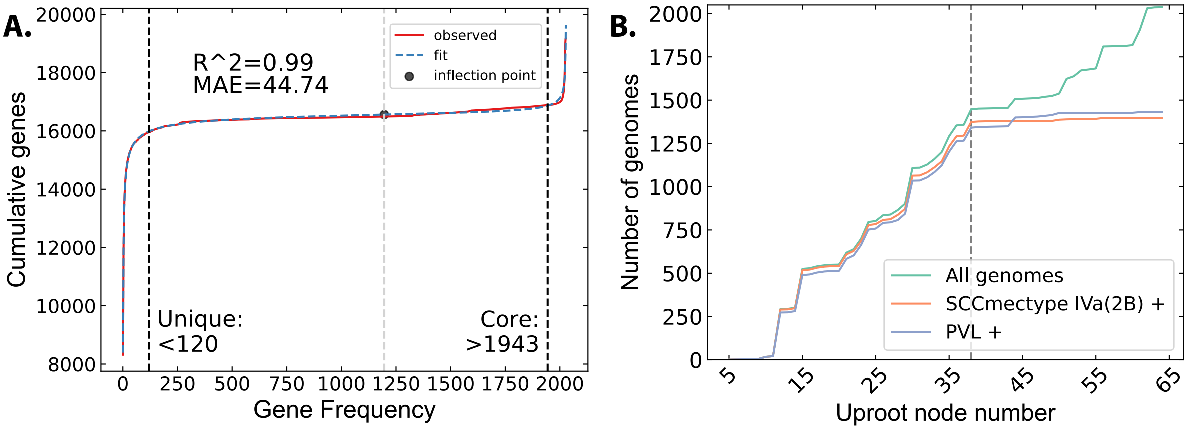


Figure C.1: (a) Cumulative distribution of unique genes used to fit the pangenomic parameters. The core and unique genes threshold were calculated at 90% of the distance from the inflection point (black dot) of the curve. (b) SCCMec and PVL distribution in the CC8 tree as it is traversed up from FPR3757 leaf towards the root. Starting from FPR3757 gives the same deliniation between USA300 and non-USA300 genomes as the search that starts from TCH1516.

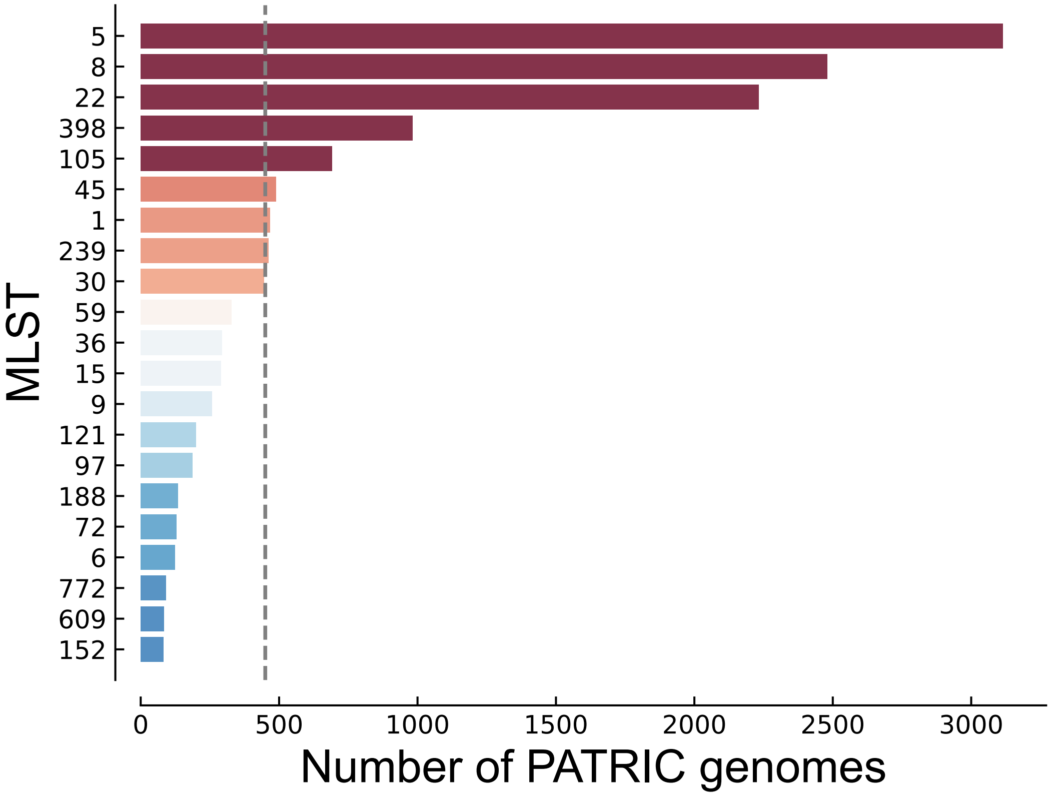


Figure C.2: *S. aureus* MLST distribution in PATRIC database

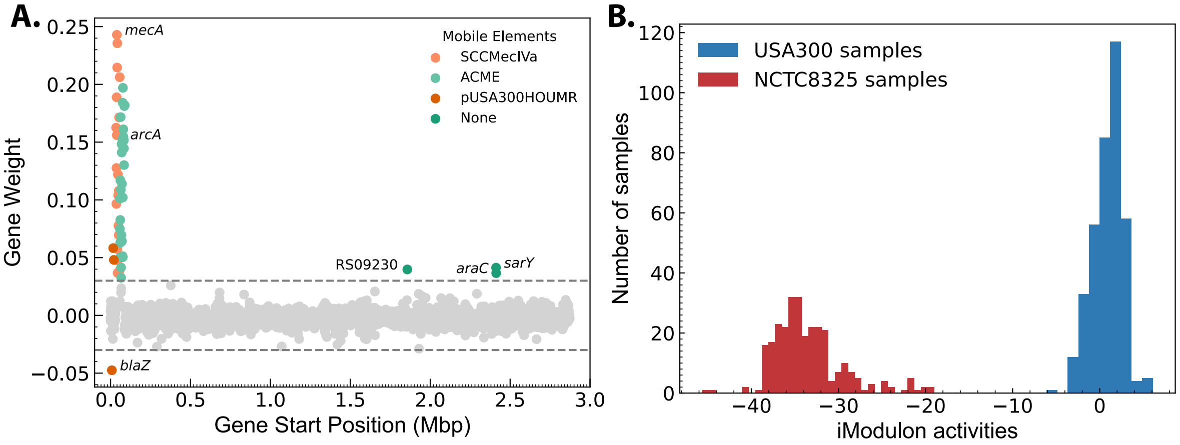


Figure C.3: SCCMec/ACME iModulons weighting and strain-specific activity. (a) Gene weighting for the iModulon primarily containing SCCMec and ACME. Genes encoding SarY and AraC family proteins were also enriched. (b) The activity of the SCCMec/ACME iModulon shows clear strain-specific separation.

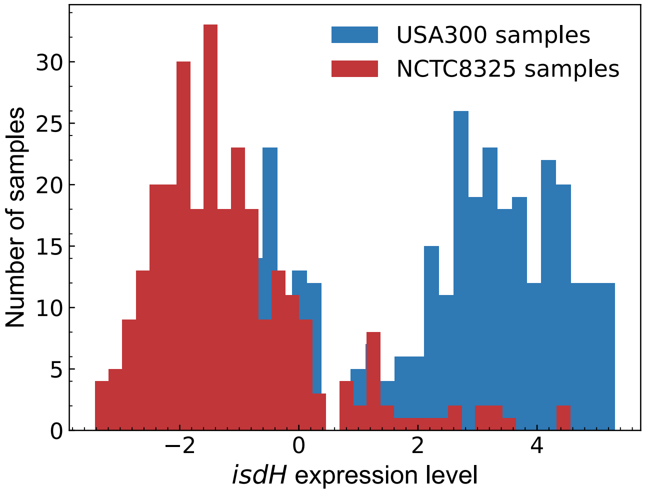


Figure C.4: *isdH* gene shows strain-specific gene expression level. The increased expression level in USA300 is in line with the deletion of Fur repressor binding site. The expression levels are log-TPM centered on TCH1516 strain grown in RPMI + 10%LB.