UNIVERSITY OF CALIFORNIA SAN DIEGO

Pan-Genome Enriched Analytics: Systems Biology Examination of the Diversity of Microbial Pathogens

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

in

Bioengineering

by

Charles Joseph Norsigian

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Nathan Lewis
Professor Victor Nizet
Professor Joseph Pogliano
Professor Bruce Wheeler
Professor Karsten Zengler

2021

The dissertation of Charles Joseph Norsigian is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents and Shannon, for their love and support that has made this work possible

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are numerous people who deserve thanks for helping me achieve this milestone. First, I would like to thank my advisor Professor Bernhard Palsson for his guidance over the past four and a half years. Dr. Palsson has cultivated a truly unique culture within the Systems Biology Research Group that offers exceptional and uncommon opportunities to its members. I am incredibly thankful for these opportunities and hope that I have successfully maximized them. It has been my privilege to be a part of the SBRG and it is an experience that I will proudly carry with me throughout my career. Dr. Palsson's openness to new ideas and trust in my ability to lead projects have enabled me to pursue and achieve an incredible amount throughout my PhD studies. His strategic thinking and scientific advice provided valuable motivation and support for this work. I also thank all of the members of my thesis committee, who have helped to guide this dissertation to its completion.

Second, I have to thank Jon Monk for being an incredible mentor throughout my time in graduate school and without whom the work presented here would not be possible. Jon and I worked together on nearly the entirety of my projects and he is responsible for helping guide me to become the scientist I am today. It has been through our work together that I have developed a passion for studying systems biology. His contributions and advice on each study and my scientific career as a whole have been immeasurable. Jon was always available to troubleshoot and consistently offered new ways to tackle any roadblocks found along the way. Further, I am grateful for his ability to remind and reinforce the value of this work throughout this process.

Outside of my primary mentors, I have to thank many of the members of the SBRG who's time in the group overlapped with mine. The culture of innovation and collaboration cherished by all group members is made possible through excellent leadership across subgroups. While my

work was not within their groups, I would like to thank Adam Feist, Dan Zielinski, Zak King, and Laurence Yang for leading groups that each contributed to the culture of the SBRG as a whole in incredibly valuable ways. I would also like to thank Marc Abrams for making the group run smoothly. I owe a large amount of thanks to my fellow lab-mates throughout my graduate school experience. First, Xin Fang and Erol Kavvas for being my first contacts with the lab for rotations and then great colleagues throughout the rest of my experience. Saugat Poudel, JC Lachance and Anand Sastry provided invaluable support and were some of my greatest friends during these years. A large thanks to all the other SBRGers it has been my pleasure to work with including Yara Seif, Patrick Phaneuf, Jared Broddrick, and many more. Additionally, I would like to thank the SBRG members of the past who laid the groundwork for building this research group and give the SBRG its exceptional institutional memory.

I also have to thank our collaborators at the Baylor College of Medicine, without whom the work in this dissertation on *C. difficile* would not have been possible. A thanks to Prof. Jennifer Spinler, Prof. Rob Britton, Dr. Heather Danhof, Dr. Firas Midani, and Colleen Brand for generating this experimental data. An extra thank you to Prof. Jennifer Spinler who's excitement and effort on these manuscripts was exceptional and her contributions greatly strengthened this work.

Prior to coming to UCSD, I had excellent mentors in University of Virginia Professors Michael Gorman, Mary Beck, and Timothy Allen, each of whom encouraged me to pursue graduate school and meaningfully invested in both my education and life. A heartfelt thank you to my high-school AP Calculus teacher, Dr. Ming-Chwan Chow, who sparked my interest in engineering and served as a foundational academic mentor who's lessons inform my approach to this day.

Last and certainly not least I have to thank all of my friends and family for all of their unconditional support and encouragement as well as good times shared through the years. The graduate school experience is marked by many peaks and valleys and I am blessed to have so many people in my life to share in the successes and rely on during the failures. Throughout this period the visits and trips shared with family and friends have been the anchors on my calendar that make the hard work all worthwhile. To Matt, Brandon, Danny, and Anthony I'm incredibly grateful to have such amazing lifelong friends. To Gus, Sasan and Arjun your friendship has greatly helped me through. I'm lucky to have grown up with a sibling who I've also called a friend throughout my life. Thanks to my sister, Meg, for always spending the time to help me in any way you can. Finally, I would like to thank the three people to whom I have dedicated this dissertation: my parents and Shannon. My parents have always encouraged my pursuits and been my greatest confidants throughout my life. I could not have done this without your love and support and you have been the greatest teachers and mentors I could ask for. You have taught me to live my life with intention, character ethics, and effort. While these words cannot fully capture my gratitude, I hope you know how proud I am to be your son. To Shannon, I can't believe that a run club trip with JC one wednesday evening after a lab-meeting ended up changing my life, but it is certainly not hyperbole to say so. You have improved my life in more ways than you know and the light of each one of my days is simply you. Thank you for your support, your companionship, and your partnership. I cannot wait to keep building our life together.

Additionally, I must thank the funding sources that supported me financially to pursue the research presented in this thesis. The National Institutes of Health (U01-AI124316) and Novo Nordisk Foundation (NNF10CC1016517) provided my support throughout graduate school.

Chapter 1, in part, is a reprint of material published in: **Norsigian CJ**, Fang X, Palsson BO, Monk JM. "Pangenome Flux Balance Analysis Toward Panphenomes." In *The Pangenome* 2020 (pp. 219-232). *Springer* The dissertation author was the primary author.

Chapter 2, in part, is a reprint of material published in: **Norsigian, Charles J.**, Erol Kavvas, Yara Seif, Bernhard O. Palsson, and Jonathan M. Monk. "iCN718, an updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE." *Frontiers in genetics* 9 (2018): 121. The dissertation author was the primary author.

Chapter 3, in part, is a reprint of material published in: **Norsigian, Charles Joseph**, Heba Attia, Richard Szubin, Aymin Yassin, Bernhard O. Palsson, Ramy K. Aziz, and Jonathan Monk. "Comparative Genome-scale Metabolic Modelling of Metallo-beta-Lactamase–producing Multidrug-resistant Klebsiella pneumoniae Clinical Isolates." *Frontiers in cellular and infection microbiology* 9 (2019): 161. The dissertation author was the primary author.

Chapter 4, in part, is a reprint of material published in: **Norsigian CJ**, Danhof HA, Brand CK, Oezguen N, Midani FS, Palsson BO, Savidge TC, Britton RA, Spinler JK, Monk JM. "Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence." *NPJ systems biology and applications.* (2020): 6. The dissertation author was the primary author.

Chapter 5, in part, is a reprint of material published in: **Norsigian, Charles J.***, Xin Fang*, Yara Seif, Jonathan M. Monk, and Bernhard O. Palsson. "A workflow for generating multi-strain genome-scale metabolic models of prokaryotes." *Nature Protocols* (2019): 1-14. The dissertation author is one of the two primary authors.

Chapter 6, in part, is a reprint of material published in: **Norsigian, Charles J.**, Neha Pusarla, John Luke McConn, James T. Yurkovich, Andreas Dräger, Bernhard O. Palsson, and

Zachary King. "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree." *Nucleic acids research* 48, no. D1 (2020): D402-D406. The dissertation author was the primary author.

Chapter 7, in part, is currently being prepared for submission for publication: **Norsigian CJ**, Danhof HA, Brand CK, Midani FS, Broddrick JT, Savidge TC, Britton RA, Palsson BO, Spinler JK, Monk JM. "Systems biology approach to functionally assess the *Clostridioides difficile* pan-genome reveals genetic diversity with discriminatory power." The dissertation author is the primary author.

VITA

2016	Bachelor of Science in Biomedical Engineering, University of Virginia
2021	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Kavvas, Erol S., Yara Seif, James T. Yurkovich, **Charles Norsigian**, Saugat Poudel, William W. Greenwald, Sankha Ghatak, Bernhard O. Palsson, and Jonathan M. Monk. "Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions." *BMC systems biology* 12, no. 1 (2018): 25.

Norsigian, Charles J., Erol Kavvas, Yara Seif, Bernhard O. Palsson, and Jonathan M. Monk. "iCN718, an updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE." *Frontiers in genetics* 9 (2018): 121.

Norsigian, Charles Joseph, Heba Attia, Richard Szubin, Aymin Yassin, Bernhard O. Palsson, Ramy K. Aziz, and Jonathan Monk. "Comparative Genome-scale Metabolic Modelling of Metallo-beta-Lactamase–producing Multidrug-resistant Klebsiella pneumoniae Clinical Isolates." *Frontiers in cellular and infection microbiology* 9 (2019): 161.

Norsigian, Charles J.*, Xin Fang*, Yara Seif, Jonathan M. Monk, and Bernhard O. Palsson. "A workflow for generating multi-strain genome-scale metabolic models of prokaryotes." *Nature Protocols* (2019): 1-14.

Norsigian, Charles J., Neha Pusarla, John Luke McConn, James T. Yurkovich, Andreas Dräger, Bernhard O. Palsson, and Zachary King. "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree." *Nucleic acids research* 48, no. D1 (2020): D402-D406.

Norsigian CJ, Fang X, Palsson BO, Monk JM. "Pangenome Flux Balance Analysis Toward Panphenomes." In *The Pangenome* 2020 (pp. 219-232). *Springer*

Jensen CS, **Norsigian CJ**, Fang X, Nielsen XC, Christensen JJ, Palsson BO, Monk JM. "Reconstruction and validation of a genome-scale metabolic model of Streptococcus oralis (iCJ415), a human commensal and opportunistic pathogen." *Frontiers in Genetics.* (2020) 11:116.

Norsigian CJ, Danhof HA, Brand CK, Oezguen N, Midani FS, Palsson BO, Savidge TC, Britton RA, Spinler JK, Monk JM. "Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence." *NPJ systems biology and applications.* (2020): 6.

Broddrick JT, Szubin R, **Norsigian CJ**, Monk JM, Palsson BO, Parenteau MN. "High-quality genome-scale models from error-prone, long-read assemblies." *Frontiers in microbiology.* (2020): 2829.

* equal contribution

ABSTRACT OF THE DISSERTATION

Pan-Genome Enriched Analytics: Systems Biology Examination of the Diversity of Microbial Pathogens

by

Charles Joseph Norsigian

Doctor of Philosophy in Bioengineering

University of California San Diego, 2021

Professor Bernhard Ø. Palsson, Chair

Technological advancements have led to an exponential increase in omics data generation. This data presents a unique big-data-to-knowledge challenge and in turn opportunity for analysis and interpretation. The construct of the pan-genome and its subsets when paired with systems biology tools, such as genome-scale models of microbial metabolism, offer a variety of means to generate meaningful predictions from genome sequence alone. The pairing of these frameworks allows for a scalable, data-driven, comparative approach to study the evolutionary trajectories of bacterial species. This dissertation focuses on the development and deployment of pan-genome

analytics tools towards the study of numerous high-threat level microbial pathogens.

Chapter 1 introduces key systems biology techniques and concepts used throughout this dissertation in particular with regard to the importance of scale of datasets.

Chapter 2 focuses on the generation of a new updated reconstruction for *Acinetobacter baumannii* and analysis of gene conservation and catabolic capabilities across the species.

Chapter 3 details comparative genome scale metabolic modeling on multidrug-resistant strains of *Klebsiella pneumoniae* and evaluates the ability of metabolic capabilities to inform on resistance profiles.

Chapter 4 describes the generation of a new reconstruction of *Clostridioides difficile* and use of this resource to evaluate the microenvironmental pressures of laboratory isolates as well as a detailed evaluation of the core-genome of the species.

Chapter 5 delineates the multi-strain reconstruction protocol used in many of the other chapters and numerous other studies providing this workflow as a resource to the research community.

Chapter 6 conducts an in-depth update to the BiGG Models knowledge base both improving the scope and diversity of content and integrating new functionalities commensurate with the directions of the field.

Chapter 7 engages in comparative analysis of *Clostridioides difficile* strains and details the development of a novel strain typing method that groups strains based on accessory genomes. These strain typings are compared in detail to the leading strain-typing schemes within the epidemiology of *C. difficile* infection and used to identify defining genetic features.

Chapter 8 provides a reflection on the state of pan-genomic applications and future directions for systems biology.

Chapter 1

Introduction

Studies of the pangenome have been empowered by an exponentially increasing amount of strain-specific genome sequencing data. With this data deluge comes a need for new tools to contextualize, analyze, and interpret such a vast amount of information. Network reconstructions, genome-scale metabolic models (GEMs), and the corresponding computational analysis frameworks such as flux balance analysis (FBA) have been proven useful toward this end. Network reconstructions can be used to interpret genomic variation not just from a single strain but for an entire species. By applying these approaches at the pangenome scale, it becomes possible to systematically evaluate phenotypic properties for an entire species thus enabling the study of diverse phenotypes directly from a pangenome. Applying insights gained from analysis of the genotype to phenotype diversity has far-ranging implications with applications ranging from human health to metabolic engineering. The future of pangenomics will include linked phenotypes analyses, thus supplementing traditional pangenomic analyses and helping to address the big-data-to-knowledge grand challenge of analyzing thousands of genomic sequences.

Conceptualizing differences between strains in a species using the construct of a

pangenome revolutionized the field of comparative genomics for bacteria [1, 2]. This framework allowed scientists to overcome problems related to species with high genomic variability and lack of a reference genome. Despite its utility, the pangenome alone cannot be used to quantify the phenotypic effects of genome variability within a species. Over the past decade, network reconstructions have become an indispensable tool in molecular systems biology because of their ability to provide a mechanistic link between experimental studies and computational analyses [3]. Thus, genome-scale network reconstructions provide an avenue for extending the power of the pangenome towards evaluating the phenotypic capabilities of a species or the panphenome. High-quality reconstructions can be expanded through bioinformatic techniques to map information from a reference strain to additional strains of the target organism.

This chapter describes how reconstructions and genome-scale models have been applied to study the pangenome by predicting all possible phenotypes for strains in a species. Using these tools, large scale genomic data sets combined with experimental phenotypes can now be integrated and queried to systematically probe the diversity of strains within a species. Genome-scale metabolic network reconstructions have been used to delineate conserved and strain-specific metabolic capabilities as well as relate differences in metabolic capabilities with lifestyle diversity across a species. This knowledge can effectively be used to define the metabolic potential of a bacterial species. In this chapter we introduce the following concepts that are fundamental to this dissertation: (1) The foundation of reconstructions and flux balance analysis; (2) The extension of these tools using a "multi-strain" approach to calculate metabolic phenotypic potential; and (3) extension of the multi-strain approach beyond metabolism.

1.1 Network Reconstructions and Flux Balance Analysis

The growing collections of sequences that have been used to study pangenomes are laden with valuable information, however, strings of nucleotide bases alone do not make this information easily accessible or immediately apparent. Thus, there is a critical need for tools that can be used to interrogate this massive amount of data to generate new knowledge. Genome-scale network reconstructions in concert with flux balance analysis (FBA) provide such a tool. This section describes the process of reconstruction as well as mathematical approaches that can be used to query and compute with reconstruction, in particular, FBA.

1.1.1 Network Reconstructions Structure Biological Knowledge

Genome-scale reconstructions are organism-specific knowledge-bases. They are built systematically using a quality controlled bottom-up workflow that incorporates genome annotation, omics data sets, and legacy knowledge. The literature detailing construction and analysis of network reconstructions is extensive [4–6]. In brief, these tools organize knowledge into a structured format linking genes, gene products and cellular components. Reconstructions can be made for several cellular processes including transcriptional regulation [7, 8], expression [9] and metabolism [10]. The reconstruction approach is iterative and thus all reconstructions are continually improving as new knowledge is generated. Thus, reconstructions serve as a valuable resource to integrate and reconcile biochemical data allowing researchers to collaborate, test, and readily share new hypotheses about functions in a target organism [11].

Reconstructions of cellular metabolism have been the most developed and extensively used type thus far [3]. Metabolic network reconstructions are comprised of all known metabolic genes, their encoded proteins and catalyzed reactions. All of this information is assembled from a range

of sources including organism specific databases, high-throughput data, and primary literature [5]. This process can also be partially automated [12, 13]. Establishing a set of the biochemical reactions that constitute a reaction network in a target organism culminates in a database of chemical equations. Reactions are then organized into pathways, pathways into subsystems, and ultimately into genome-scale networks; thus representing biological processes at multiple scales. Network reconstructions represent an organized process for genome-scale assembly of disparate information about a target organism. All this information is put into context with the annotated genome to form a coherent whole. Today, there exist collections of genome-scale reconstructions for a number of target organisms across the tree of life [11, 14]. For example, as of 2018 there are 178 available, curated reconstructions spanning the tree of life. While this coverage is impressive, several other phyla remain devoid of any reconstruction initiative. To fully extend the study of pan-phenomes to all sequenced organisms, new reconstruction efforts must be initiated [11].

1.1.2 Flux Balance Analysis Enables Computation of Phenotype from Genotype

Reconstructions alone are static, and unable to be used for predictions. A major value of the metabolic reconstructions emerges when they are converted into a mathematical format, thus becoming amenable to computational interrogation using a variety of computational methods [15, 16]. This conversion translates a reconstructed network into a chemically accurate mathematical format that becomes the basis for a genome-scale model (GEM). This conversion requires the mathematical representation of metabolic reactions. The core feature of this representation is tabulation, in the form of a numerical matrix, of the stoichiometric coefficients of each reaction. These stoichiometries impose systemic constraints on the flow of metabolites through the network

represented as balances or inequalities for bounds [17]. Further constraints can be added to a network such as thermodynamic reversibility constraints and limitations to nutrient uptake or byproduct secretion. Computationally predicted network states that are consistent with all imposed constraints are thus candidate physiological states of the target organisms under a given defined condition.

Flux balance analysis (FBA) can be applied to these models for prediction of an organism's phenotype. This mathematical approach for analyzing the flow of metabolites through a metabolic network is the original constraints based method [15]. This approach relies on an assumption of steady-state growth and mass balance. FBA uses the stated objective (for example, biomass production, e.g. growth) to find the solution(s) using linear programming that optimize the objective function [4]. The inner workings of a GEM are readily understood conceptually. In a given, defined environment, GEMs can be used to compute network outputs based on defined inputs. FBA can computationally trace a fully balanced path through the reactome from the available nutrients to the prerequisite output metabolite. Using FBA, a GEM can compute the balanced use of the reactome to produce all the prerequisite metabolites for growth simultaneously, and does so in the correct relative amounts while accounting for all the energetic, redox, and chemical interactions that must balance to enable such biomass synthesis [4].

Using this technique, a variety of phenotypes such as the effect of gene knockouts, metabolite secretion and growth capabilities on different substrates can be predicted rapidly and compared to experimental results to verify their accuracy [18]. Some of the best models have accuracies greater than 90% in agreement with experimental data [19, 20]. In this way GEMs provide a way to bridge the genotype to phenotype gap by providing a robust platform for analyzing the integrated mechanisms of gene products to produce unique phenotypic states. The utility of a

highly-curated GEM and the corresponding computational analyses is increased by the format's scalability. Through this methodology, phenotypes for the plethora of sequenced strains within a species become readily calculatable. In the next section we will highlight how high-quality reconstructions for a single strain can be extrapolated onto several strains of the same species to study the phenotypic potential of the pangenome and to gain insight into strain-specific metabolic capabilities.

1.2 The Multi-Strain Approach: Extending Genome-Scale Models to Robustly Explore the Pangenome Phenotypic Space

Once a high-quality reconstruction and genome-scale model exist, its contents (e.g. genes, metabolites, and reactions) can be mapped onto other, closely related strains in a species. Following this multi-strain approach, tools from comparative genomics [21] can be integrated with genome-scale modelling to identify genetic determinants underlying variability at the phenotypic level. Such a task is crucial to understand the evolutionary trajectories of a bacterial species. Recently, genome-scale metabolic models of different strains have been assembled to highlight the intra-species diversity at the metabolic level. Strain-specific metabolic capabilities and auxotrophies can be predicted and used to study capabilities related to the lifestyle diversity of a bacterial species. This approach is scalable to the pangenome level and in turn enables pan-phenome analysis, thus empowering species-wide comparative systems biology. This multi-strain approach has been applied to several species in a variety of studies and we provide a brief overview of the key insights here. Further the process is formalized and described in detail within Chapter 5.

1.2.1 Genesis of the Multi-Strain Approach: Studying *Escherichia coli*

The first instance of the multi-strain approach as described here was executed by Monk et al. where the authors leveraged a curated genome-scale model of *E. coli* K-12 MG1655 that has been continually updated over 15 years to construct genome-scale models of 55 other fully sequenced *E. coli* strains [22]. Using FBA on all 55 of these models the authors were able to extensively investigate the predicted metabolic capabilities of all the strains. The authors delineated strain-specific auxotrophies and substrate preferences amongst the set of strains. It is important to note that these predictions and insights were gained from sequence alone. Further, this study demonstrated the possibility of applying this approach to understand cases of patho-adaptation to a given environment and evaluate a given strain's infectious niche.

Further work scaled up the effort to include 1200 strains of *E. coli* and demonstrated a large amount of variability within the species both in gene content and consequent variability of gene products [19]. It also utilized the differences across the 1200 strains to construct a robust classification tree for determination between extra-intestinal and intra-intestinal pathogens using predicted metabolic phenotypes. This type of classification schema opens the door to investigating how strain-specific traits impact the microbiome. An in-depth example of such analyses came in a study by Fang et al into the metabolic capabilities of inflammatory bowel disease (IBD) associated *E. coli* strains in the B2 clade [23]. The authors found these strains have advantages in catabolizing sugars derived from mucus glycans. The interesting and novel outcomes of these *E. coli* studies clearly demonstrated the value of the approach, and the natural next step was to apply the methodology to other species.

1.2.2 Expanding the Reach of Multi-Strain Approach Across the Phylogenetic Tree

Numerous studies followed the first *E. coli* studies that focused on various organisms. Fouts et al applied the multi-strain approach, broadened to examine various species of *Leptospira* known to have ranging levels of pathogenicity [24]. They demonstrated that the ability to synthesize vitamin B12 is limited to pathogenic species of *Leptospira* and may give them a survival advantage in a human host where B12 is sequestered by the body. This valuable distinguishing metabolic capability was captured by being able to leverage the base reconstruction across multiple species in the genus.

In 2016 Bosi et al applied the workflow to 64 strains of *Staphylococcus aureus*. Beyond reconstructing metabolic capabilities they extended the approach to identify virulence factors in the set of 64 strains [25]. By using a combination of predicted metabolic capabilities linked to virulence factors, they were able to stratify the strains by host-type. This study added an additional layer to the promise of the multi-strain approach by showing that metabolic capabilities could be analyzed in concert with other components of the pangenome, namely virulence factors (toxins, adhesins, etc.), and that this combination held predictive power about a strain's host. This study also included explicit calculation of the core and pangenome content of *S. aureus*, a metric of genomic diversity amongst strains in a species.

The multi-strain approach has also been applied to other pathogens such as *Acinetobacter baumannii* and *Salmonella*. The study on *A. baumannii* is detailed within Chapter 2. Seif et al built strain-specific models for 450 *Salmonella* strains from various serovars to show that metabolic capabilities can be used to distinguish these serovars [26]. This study indicates that host-range may be limited by metabolic capabilities of different strains.

1.2.3 Extending the Multi-Strain Approach to Investigate Additional Biological Qualities

The multi-strain framework provides an inherently efficient means of interrogating the properties of many strains and a few studies have utilized this organizational efficiency to gain insight into properties outside of direct metabolic capabilities. For example, Choudhary et al examined the agr type of 400 *S. aureus* strains to examine the structure of genes within the genome [27]. The authors found that genomic virulence factor profiles are highly correlated with agr type. They also identified that divergence in histidine kinase protein confers signal specificity with clear differences in protein structural properties based on agr types. Another example of additional properties is the investigation of reactive oxygen species (ROS) tolerance. By leveraging the multi-strain approach in conjunction with 3D structures Mih et al was able to simulate ROS production levels to demonstrate that antioxidant properties are exhibited in the structural proteome (Mih et al. 2018). A third example was conducted by Kavvas et al, who took a deeper level of resolution within the genome by looking at the unique alleles present within *Mycobacterium tuberculosis* genomes [28]. Through machine learning techniques on the pangenome they were able to associate certain alleles potentially responsible for antimicrobial resistance. The results hint at metabolic rewiring at the allelic level required for adaptation to antibiotic resistance.

1.3 Moving Beyond Metabolism: Multi-Scale Approaches to Species Diversity

This chapter details a computational approach (reconstruction and FBA) to systematically calculate metabolic phenotypes for multiple strains in a species. Beyond calculation of metabolic phenotypes, new techniques, both experimental and computational, offer exciting new avenues for research into the pangenome. These approaches can be applied at multiple different scales. At the lowest level, single nucleotide variants (SNV) can be compared across strains using sequence mapping toolkits like breseq and gatk [29, 30]. These techniques can be scaled up from single base changes to full gene sequences to compare orthologous ORFs across genomes by comparing sequence-specific alleles across strains in a species or the alleleome. As described here, the presence/absence of given enzyme-encoding metabolic genes can be used to build strain-specific metabolic reconstructions that compute metabolic phenotypes. While most of the applications described here are applied to pathogens with relevance to human health, it is important to note that the pangenome can also be studied for use in metabolic engineering applications. For example, the pangenome can be mined to search for enzymes of interest to industrial microbiology [31].

In the future, processes beyond metabolism will also be reconstructed allowing for full panphenome calculations. For example, reconstructions of expression mechanisms already exist [9] and have been integrated with models of metabolism (ME models) [32].These models account for the transcription and translation processes and molecular constituents required to express enzymes catalyzing metabolic reactions in the metabolic network. In the future, multiple ME models of strains in a species will further expand the scope of computation possible on contents

of the pangenome.

Beyond metabolism and expression, regulatory networks are another aspect of the pangenome that differ between strains and have been reconstructed for individual strains [7, 8]. Understanding how certain strains regulate the same set of genes (core-genome) as well as diverse sets of genes will further expand our understanding of the structure and function of the pangenome. A small scale study of seven *E. coli* strains and their RNA-seq expression profiles in aerobic and anaerobic environments showed remarkably different expression levels even for shared genes of the core-genome [33]. Studying differentially expressed genes and the transcription factors known to regulate them may lead to discovery of alternative regulatory strategies between strains in a species.

Just as sequences databases have grown tremendously in recent years, 3D crystal structures for the encoded genes have also grown dramatically [34]. The protein data bank [35] (PDB) is a repository of protein structures and these structures can now be integrated with genome-scale models (GEM-PRO) [36]. Building multi-strain models with associated protein structures is another way to compare strains across a species. Using these tools, sequence diversity can be examined at the 3D level to see how mutations line up in 3D space, a level of analysis not possible at the sequence level. Furthermore, mutations in specific regions of the protein can be tabulated and compared across strains [37].

Finally, a multi-strain approach should prove useful for studies of the microbiome. Multiple genome-scale models for species found in the microbiome already exist [38]), and GEM studies were proven effective in studying the impact of diet [39] and interactions between microbes [40]. Expanding the multi-strain approach to study diverse strains in these species may lead to deeper level understanding of the gut-microbiome composition. Indeed, strain-level metagenomics is

coming [41] and expanding study of the pangenome to the microbiome will have fruitful applications in the near future.

We must also acknowledge some caveats and risks to the multi-strain approach. First, all of these approaches require high-quality sequence data connected to high-quality, QC/QA data generation. The success of reliable and maximally effective future pan-phenomics rests on ensuring this quality. There must be a continued effort to ensure that sequencing projects are of quality not only quantity. Additionally, an interesting question pertaining to the concept of closed pangenomes is how will the law of diminishing returns be exhibited in these sequence deposits. Will a point be reached where additional sequences provide no novel information? Further, the vision of the pan-phenome and its implications to understanding how microbial pathogens impact human health will rely on both the availability of metadata and the deposition of strains. Metadata on these strains will only deepen the possible questions to be asked of both pangenomes and panphenomes. A centralized repository of strains will also greatly expedite the experimental verification needed for such large computational predictions. The future of the panphenome is apparent and with it further explanations at the center of biological causality.

1.4 Perspective

Significant advancements in DNA sequencing technology have led to an exponential increase in the number of sequenced strains. This creates a need for new ways to integrate and analyze this ever-increasing amount of sequence information. This need will only intensify as the number of sequenced strains within a species continues to grow exponentially. This chapter demonstrates how the pangenome is evolving from a theoretical concept to a queryable construct.

In this chapter, we describe the foundational aspects of GEMs and FBA and their use to

predict phenotypic states for multiple strains in a species. The multi-strain approach has proven useful in extending this utility in a number of studies providing evolutionary insights as well as practical applications. As the library of available sequences continues to grow, the possibility of scaling these techniques to the level of the pangenome across the tree of life is becoming a reality. The ability to systematically characterize an entire species' phenotypic capabilities will enhance the depth of pangenome analysis possible and pull valuable information inherent to genome sequences to the forefront. The linkages and distinct features at the pangenome scale for a species offer obvious value for future knowledge generation, especially pertaining to human health and disease. Further, the future potential applications outlined here such as inclusion of expression, regulation, and structures into these workflows will only further advance the scope of genome-scale science. Genome sequences are laden with critical information and the tools/workflows described in this dissertation provide a means for extracting this information into actionable knowledge.

Each chapter in this dissertation details a systems biology approach towards analyzing a pathogenic species or details the development of computational techniques and resources critical to these approaches. Taken together, the results presented advance both the knowledge for each pathogen of interest as well as the toolkit available for pangenome analytics.

Acknowledgements

Chapter 1, in part, is a reprint of material published in: **Norsigian CJ**, Fang X, Palsson BO, Monk JM. "Pangenome Flux Balance Analysis Toward Panphenomes." In *The Pangenome* 2020 (pp. 219-232). *Springer* The dissertation author was the primary author.

1.5 References

1. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Scott Durkin, A., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Ros, I. M. y., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13950–13955 (Sept. 2005).

2. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. en. *Curr. Opin. Genet. Dev.* **15,** 589–594 (Dec. 2005).

3. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nat. Rev. Genet.* **15,** 107–120 (Feb. 2014).

4. 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).

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

6. Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Reconstruction of biochemical networks in microorganisms. *Nat. Rev.* (2008).

7. Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R. & Palsson, B. O. Matrix formalism to describe functional states of transcriptional regulatory systems. en. *PLoS Comput. Biol.* **2,** e101 (Aug. 2006).

8. Gianchandani, E. P., Joyce, A. R., Palsson, B. Ø. & Papin, J. A. Functional states of the genome-scale Escherichia coli transcriptional regulatory system. en. *PLoS Comput. Biol.* **5,** e1000403 (June 2009).

9. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. Ø. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* **5,** e1000312 (2009).

10. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. en. *Nat. Rev. Microbiol.* **7,** 129–143 (Feb. 2009).

11. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. en. *Nat. Biotechnol.* **32,** 447–452 (May 2014).

12. Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. & Stevens, R. L. High-throughput generation, optimization and analysis of genome-scale metabolic models. en. *Nat. Biotechnol.* **28,** 977–982 (Sept. 2010).

13. Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I. & Nielsen, J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. en. *PLoS Comput. Biol.* **9,** e1002980 (Mar. 2013).

14. Oberhardt, M. A., Puchałka, J., Martins dos Santos, V. A. P. & Papin, J. A. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. en. *PLoS Comput. Biol.* **7,** e1001116 (Mar. 2011).

15. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? en. *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

16. 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).

17. Reed, J. L. Shrinking the metabolic solution space using experimental datasets. en. *PLoS Comput. Biol.* **8,** e1002662 (Aug. 2012).

18. Monk, J. & Palsson, B. O. Genetics. Predicting microbial growth. en. *Science* **344,** 1448–1449 (June 2014).

19. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

20. Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., Prlić, A., Sastry, A., Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M. T., Nielsen, J., Thiele, I. & Palsson, B. O. Recon3D enables a three-dimensional view of gene variation in human metabolism. en. *Nat. Biotechnol.* **36,** 272–281 (Mar. 2018).

21. Monk, J. & Bosi, E. in *Metabolic Network Reconstruction and Modeling: Methods and Protocols* (ed Fondi, M.) 151–175 (Springer New York, New York, NY, 2018).

22. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

23. Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L. & Palsson, B. O. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. en. *BMC Syst. Biol.* **12,** 66 (June 2018).

24. Fouts, D. E., Matthias, M. A., Adhikarla, H., Adler, B., Amorim-Santos, L., Berg, D. E., Bulach, D., Buschiazzo, A., Chang, Y.-F., Galloway, R. L., Haake, D. A., Haft, D. H., Hartskeerl, R., Ko, A. I., Levett, P. N., Matsunaga, J., Mechaly, A. E., Monk, J. M., Nascimento, A. L. T., Nelson, K. E., Palsson, B., Peacock, S. J., Picardeau, M., Ricaldi, J. N., Thaipandungpanit, J., Wunder Jr, E. A., Yang, X. F., Zhang, J.-J. & Vinetz, J. M. What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira. en. *PLoS Negl. Trop. Dis.* **10,** e0004403 (Feb. 2016).

25. 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. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

26. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

27. 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).

28. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nat. Commun.* **9,** 4306 (Oct. 2018).

29. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. en. *Methods Mol. Biol.* **1151,** 165–188 (2014).

30. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. en. *Genome Res.* **20,** 1297–1303 (Sept. 2010).

31. Moscatello, N. & Pfeifer, B. A. Constraint-based metabolic targets for the improved production of heterologous compounds across molecular classification. *AIChE J.* **9,** 293 (July 2018).

32. O'brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9,** 693 (2013).

33. Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J. & Feist, A. M. Multi-omics Quantification of Species Variation of Es-

cherichia coli Links Molecular Features with Strain Phenotypes. en. *Cell Syst* **3,** 238–251.e12 (Sept. 2016).

34. Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E. J., Bliven, S. E., Chen, K., Chang, R. L., Bourne, P. E. & Palsson, B. O. Systems biology of the structural proteome. en. *BMC Syst. Biol.* **10,** 26 (Mar. 2016).

35. Berman, H. M., Westbrook, J., Feng, Z., *et al.* The protein data bank. *Nucleic acids* (2000).

36. Chang, R. L., Andrews, K., Kim, D., Li, Z., Godzik, A., *et al.* Structural Systems Biology Evaluation of Metabolic Thermotolerance.

37. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. ssbio: a Python framework for structural systems biology. en. *Bioinformatics* **34,** 2155–2157 (June 2018).

38. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T. & Thiele, I. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. en. *Nat. Biotechnol.* **35,** 81–89 (Jan. 2017).

39. Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters, T., Juste, C., Rizkalla, S., Chilloux, J., Hoyles, L., Nicholson, J. K., MICRO-Obes Consortium, Dore, J., Dumas, M. E., Clement, K., Bäckhed, F. & Nielsen, J. Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. en. *Cell Metab.* **22,** 320–331 (Aug. 2015).

40. Shoaie, S., Karlsson, F., Mardinoglu, A., Nookaew, I., Bordel, S. & Nielsen, J. Understanding the interactions between bacteria in the human gut through metabolic modeling. en. *Sci. Rep.* **3,** 2532 (2013).

41. Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L. & Segata, N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. en. *Nat. Methods* **13,** 435–438 (May 2016).

Chapter 2

iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of *Acinetobacter baumannii* AYE

2.1 Abstract

Acinetobacter baumannii has become an urgent clinical threat due to the recent emergence of multi-drug resistant strains. There is thus a significant need to discover new therapeutic targets in this organism. One means for doing so is through the use of high-quality genome-scale reconstructions. Well-curated and accurate genome-scale models (GEMs) of *A. baumannii* would be useful for improving treatment options. We present an updated and improved genome-scale

reconstruction of *A. baumannii* AYE, named iCN718, that improves and standardizes previous *A. baumannii* AYE reconstructions. iCN718 has 80% accuracy for predicting gene essentiality data and additionally can predict large-scale phenotypic data with as much as 89% accuracy, a new capability for an *A. baumannii* reconstruction. We further demonstrate that iCN718 can be used to analyze conserved metabolic functions in the *A. baumannii* core genome and to build strain-specific GEMs of 74 other *A. baumannii* strains from genome sequence alone. iCN718 will serve as a resource to integrate and synthesize new experimental data being generated for this urgent threat pathogen.

2.2 Introduction

Acinetobacter baumannii has recently emerged as a deadly nosocomial threat with rising rates of both infection and antibiotic resistance. Reports using data from hospital-based surveillance studies as well as from the Infectious Diseases Society of America have begun to refer to a dangerous group of nosocomial pathogens, including *A. baumannii*, as "ESKAPE pathogens" [1]. *A. baumannii* in particular is known for its highly persistent and opportunistic nature, most often resulting in hospital-acquired pneumonia while also having the ability to infect various other tissues [2]. Organisms of the genus *Acinetobacter* inhabit a wide variety of environments, ranging from humans to water and soil [3]. These diverse environmental niches are reflected in the genomic content of the organisms as well as their metabolic capabilities. *Acinetobacter* are Gram-negative, aerobic, and non-motile. Pathogenic *A. baumannii* antibiotic resistance has risen from a susceptible level in the 1960s to extended and pan-drug resistant today [4]. As such, the need for new treatment targets and strategies is dire.

Genome-scale models (GEMs) of metabolism have been used to discover new drug targets

[5] and pursue novel treatment options. Genome-scale metabolic reconstructions offer an established framework for systems-level analyses of an organism's metabolism [6]. GEMs provide a formal way to link genotype to phenotype and mechanistically analyze the metabolic capabilities of organisms. A previous reconstruction of the metabolic network of *A. baumannii* AYE was undertaken and produced: AbyMBEL891 [5]. This reconstruction provided a valuable starting point for the progress and use of GEMs to study the pathogenic nature of *A. baumannii*. However, one issue that has limited the use of this and other reconstructions is the lack of standardization in identifiers for metabolites and reactions [7]. Since the publication of AbyMBEL891 in 2010, numerous studies have produced new data [8–10] that provide an opportunity to update this *A. baumannii* reconstruction, allowing for more accurate representations of its physiology. One such study was a high-quality reconstruction of *A. baumannii* ATCC 19606, iLP844, that served as a valuable resource for model improvements [10]. Furthermore, given that *Acinetobacter* is known to populate a diverse array of environments, particularly hospitals, it is likely that diverse metabolic capabilities may be present throughout the different strains in this species.

We present iCN718, a new and updated GEM of *A. baumannii* AYE. This reconstruction utilizes AbyMBEL891 as a foundation. We validated our model by comparing phenotypic predictions made by iCN718 to those made by AbyMBEL891. We extended our analysis to additional datasets published after AbyMBEL891. We assessed iCN718 on its ability to predict both gene essentiality and to recapitulate experimental growth capabilities. We then utilize this reconstruction to create draft models of 74 other *A. baumannii* strains from their sequence data alone. We leverage the reconstruction to produce draft models to gain insight into these other strains and the species as a whole. Thus, iCN718 offers a framework for sequence-to-model comparisons. Our updated model of *A. baumannii* will provide new opportunities to advance the

understanding of pathogenic microbes and their interactions with human hosts.

2.3 Results and Discussion

2.3.1 Workflow for Network Reconstruction

We began the metabolic network reconstruction process by updating AbyMBEL891. We found that the AbyMBEL891 reconstruction could be updated and improved in three main areas: (1) standardization of reaction and metabolite identifiers to increase the tractability of the network, (2) mass and charge balance metabolic reactions, and (3) transport processes. Before updating and improving the reconstruction, we recognized that it was necessary to translate AbyMBEL891 into a format that could be more readily analyzed. We obtained a draft reconstruction of *A. baumannii* AYE using the ModelSeed database [11]. We then cross-referenced draft reconstruction reactions against AbyMBEL891 and utilized additional databases to map all reactions and metabolites to the standardized BiGG format [12]. Additionally, we added the curated gene product rules (GPRs) from AbyMBEL891 into iCN718 to improve ease of simulation [10]. The resulting model was then continually and iteratively improved through manual curation of new organism knowledge in the literature published since the release of AbyMBEL891 (See section "Materials and Methods" and Figure 2.1).

iCN718 comprises 718 genes, 1016 reactions, and 890 metabolites compared to the 650 genes, 891 reactions, and 770 metabolites in AbyMBEL891. The majority of the difference in reactions included arises from the inclusion of exchange reactions in iCN718 as well as revamping the transport reactions. The reversibility of reactions within iCN718 was referenced against the reversibility of corresponding reactions in a recently published model of *A. baumannii* ATCC

19606, iLP844 [10]. In some cases, reaction reversibility was changed to reflect the state in iLP844. Reversibility was corroborated with iLP844 for a set of about 50 reactions and edited accordingly. iLP844 was also used to identify GPRs for transport reactions present in both models, leading to the inclusion of 66 new genes in iCN718. Further, new reactions that were missing in the original reconstruction were added in peptidoglycan biosynthesis, propanoate metabolism, and glycolate catabolism. The end product of iCN718 is a reconstruction of *A. baumannii* AYE that rectifies issues with AbyMBEL891 regarding identifiers, reversibility of reactions, transport/exchange reactions, and mass/charge balancing. Well-curated identifiers were added for every reaction in the network. Thus, iCN718 provides an improved knowledge-base for the study of *A. baumannii*.

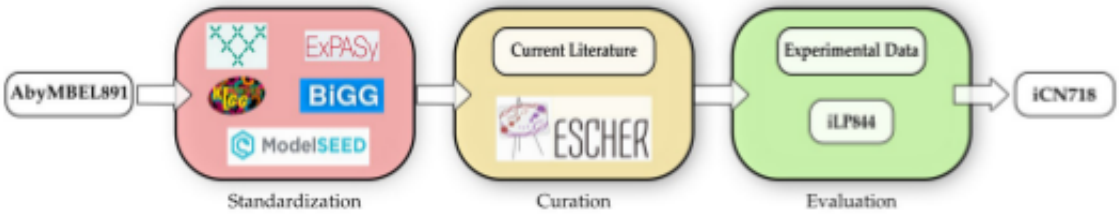

Figure 2.1: Workflow of the reconstruction process. The starting reconstruction, AbyMBEL891, was cross referenced against a draft model generated utilizing ModelSEED [11]. Next, the reconstruction was standardized using various databases mapped to standard BIGGs IDs. This process was followed by manual curation based on current literature on the organism, aided by the use of ESCHER to visualize pathways throughout the process. Finally, the model was evaluated against experimental datasets and compared to iLP844 a model of *Acinetobacter baumannii* ATCC 19606 to further improve the reconstruction. The model was iteratively evaluated against gene essentiality and phenotypic datasets to improve the reconstruction accuracy.

After completing the reconstruction of iCN718, we calculated the metabolite connectivity to evaluate the network structure for both iCN718 and AbyMBEL891 [13]. Metabolite connectivity refers to the number of reactions in which a metabolite participates. Given that metabolites are the nodes of the network connected by reactions, this metric reveals the connectivity of a metabolic network. We compared the metabolite connectivities of iCN718 and

AbyMBEL891 (Supplementary Figure A.1) and found that overall, the networks were comparable, but these plots do not visualize dead-end metabolites (i.e., metabolites with a connectivity of one). iCN718 has four dead-end metabolites whereas AbyMBEL891 has 145 dead-end metabolites, demonstrating that iCN718 is more highly connected overall. The increase in connectivity is a result of converting to BiGG standard identifiers which improves the regularity of the network.

2.3.2 Functional Evaluation of iCN718

Our first functional evaluation of iCN718 consisted of analyzing its accuracy in predicting gene essentiality for three datasets (Figure 2.2). The most comprehensive essentiality dataset available was used [9]. This complete TN-seq essentiality dataset was conducted with *A. baumannii* AB5075 and is particularly valuable because it is of genome scale and every gene in iCN718 has an ortholog. iCN718 was able to achieve 80.22% accuracy (Figure 2.2). Unfortunately, given the lack of GPRs in AbyMBEL891, we were unable to analyze its performance on this dataset. We also evaluated iCN718's performance on the two datasets originally used to validate AbyMBEL891. The first was an insertional mutagenesis dataset with *A. baumannii* ATCC 19606 by Dorsey et al. [14] on a set of 14 mutants. We repeated the same knockouts in silico as done in the original experiment and found that iCN718 was able to correctly predict 100% (14/14) of the mutant cases as did AbyMBEL891. The obvious limitation of this dataset is that it is on such a small scale. The second dataset used to validate AbyMBEL891, by de Berardinis et al. [15], was a complete, genome-scale set of single-gene deletions in *Acinetobacter* baylyi ADP1. iCN718 fell short in predictive ability on this dataset compared to AbyMBEL891 (Figure 2.2), with 68% and 72% accuracy, respectively.

The higher predictive accuracy on the Gallagher dataset compared to the de Beradinis

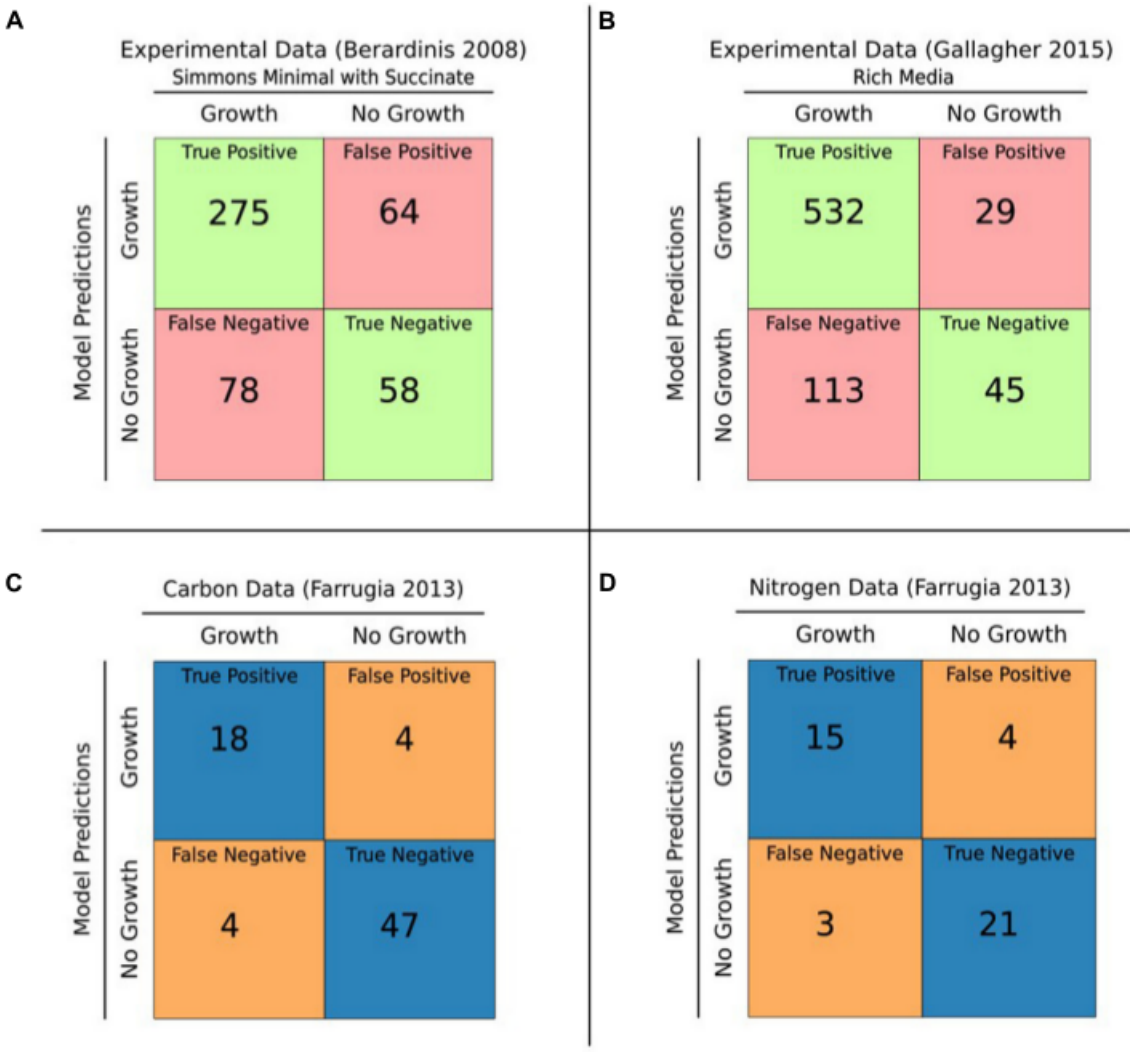

Figure 2.2: Gene essentiality and growth predictions (A) iCN718 was used to predict gene essentiality. The results were compared to the de Berardinis et al. [15] experimental dataset with 68% accuracy. (B) iCN718 predicted gene essentiality results compared with the Gallagher et al. [9] dataset exhibited 80% accuracy. It is worth noting that the Berardinis dataset was of Acinetobacter baylyi ADP1 and therefore not every gene in iCN718 had an orthologous gene in the essentiality dataset. Green represents correct predictions, red represents incorrect predictions. The Gallagher dataset is from *Acinetobacter baumannii* strain AB5075 of which there is an ortholog for every gene within iCN718. Model-predicted ability to catabolize various sole carbon (C) and sole nitrogen (D) sources compared to the Farrugia et al. [8] Biolog Phenotypic Array data for *Acinetobacter baumannii* AYE exhibited 89% and 84% accuracy, respectively. Blue represents correct predictions, orange represents incorrect predictions. Only compounds readily mapped to model metabolites were included from the Biolog data.

dataset is encouraging because strain AB5075 is a clinical isolate like AYE whereas *A. baylyi* ADP1 is a soil strain. The disparity in genomic content between *A. baumannii* AYE and *A. baylyi* ADP1 is evident in the limited number of genes in iCN718 that have an ortholog. Despite the limitations of the original two datasets, whether it be scale or lack of similarity, it was important to test iCN718's ability to recapitulate the capabilities of AbyMBEL891. Overall, iCN718 performed the same as AbyMBEL891 on the datasets originally used for validation. Further, there is more agreement of genes with a dataset on a strain that is closer to the target of the reconstruction. It is reasonable to conclude from these gene-essentiality results that at a minimum, iCN718 performs in line with AbyMBEL891 in regard to gene essentiality and more likely is superior in predictive capability. An obvious avenue for further improvement of the reconstruction would be to develop a gene essentiality dataset for strain AYE.

We further extended our assessment of iCN718 to large-scale phenotypic data. By utilizing the Biolog Phenotype Microarray data published by Farrugia et al. [8], we were able to iteratively improve iCN718 through manual curation for discrepancies. The model had encouraging agreement at the end of this process for sole carbon and nitrogen sources readily tractable to the model (116 total; Figure 2.2). Growth rates were calculated in Simmons' Minimal Medium and iteratively investigated for each carbon or nitrogen source in the microarray wet lab experiment. The model result of growth or no growth determined by optimizing for the biomass function was compared to the data from the microarray (Supplementary Tables A.1 A.2). For the carbon sources tested on the microarray plate, 73 metabolites were analyzed and showed that iCN718 has 89.1% agreement with the experimental data. Likewise, for nitrogen sources, 43 metabolites were screened with 83.7% agreement. Importantly, out of all the datasets used for validation of the reconstruction, this microarray data was the only set executed with the strain

of interest, *A. baumannii* AYE. Therefore, this dataset was particularly valuable for insight into the capabilities of this specific strain.

We have demonstrated that iCN718 performs as well as AbyMBEL891 on datasets originally used to validate AbyMBEL891. We note that these datasets suffer from limitations in that they are either not genome scale or are not of an ideally similar species to the strain of interest. To expand the validation of iCN718 and address these limitations, we analyzed a genome-scale set of gene essentiality data of another *A. baumannii* clinical strain and found a reasonably high level of agreement. Further we analyzed iCN718's agreement with phenotypic microarray experiments conducted with strain AYE. iCN718's ability to capture this growth behavior is a major improvement over AbyMBEL891, which fails to simulate on the minimal media conditions corresponding to these experiments. Overall, we showed that iCN718 maintains comparable performance on the original datasets used for validation, has a higher agreement with gene essentiality data for a more closely related strain, and is able to correctly predict phenotypic growth experiments (Figure 2.3). We used the model to perform synthetic lethals analysis to generate new predictions. Briefly this resulted in 49 synthetic lethal gene pairs that include 62 unique genes. These genes correspond to reactions involved in fatty acid metabolism, purine metabolism, glycine/serine/threonine metabolism, phenylalanine/tyrosine/tryptophan biosynthesis, TCA cycle, lysine degradation, glycerophospholipid metabolism, glycolysis, pyrimidine metabolism, nicotinate/nicotinamide metabolism, riboflavin metabolism, pentose phosphate pathway, cysteine metabolism, and methionine metabolism. Synthetic lethal gene pairs are reported in Supplementary Table A.4.

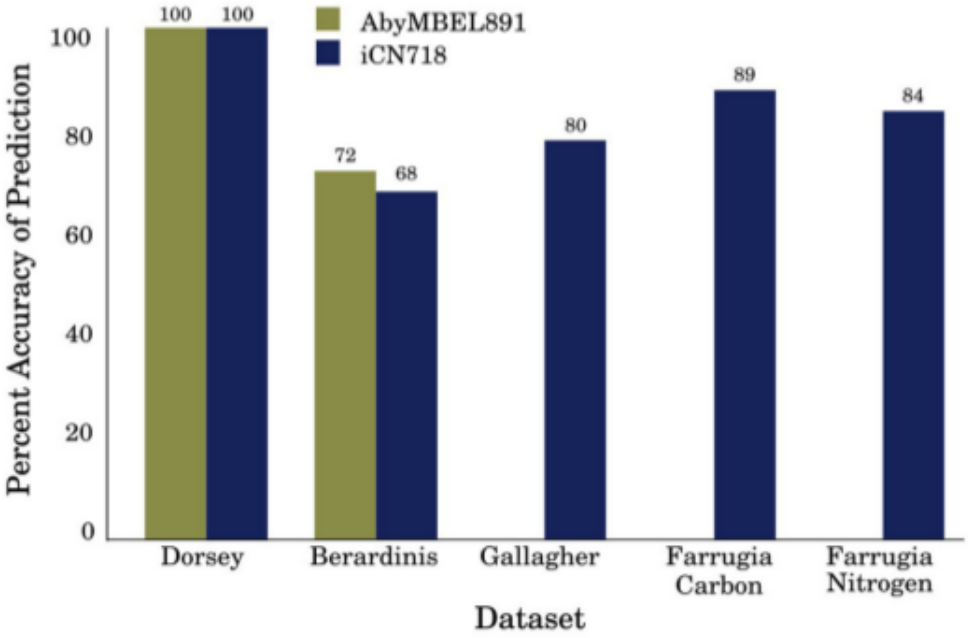

Figure 2.3: Summary of AbyMBEL891 and iCN718 Performance. Overall performance of iCN718 compared to a previous *Acinetobacter baumannii* AYE reconstruction (AbyMBEL891). Both models perform similarly on the datasets originally used to validate AbyMBEL891; however, the ability to simulate sole carbon and nitrogen sources in minimal media is exclusive to iCN718. AbyMBEL891 could not be simulated with the Gallagher dataset and was incapable of growth in the conditions of the Farrugia dataset.

2.3.3 Pan-Genome Analysis of *A. baumannii* using iCN718

A GEM can be used to investigate the capabilities of organisms across multiple strains. We applied these principles using iCN718 to explore the different genotypes and phenotypes within the *A. baumannii* species. There are 75 full complete sequences of *A. baumannii* available on the PATRIC database [16]; these range from a wide variety of isolation countries and are largely isolates from a clinical/human setting (See Supplementary Table A.5). We collected the annotated open reading frames (ORFs) from each of these genomes and used CD-HIT [17] to assign their coding sequences into clusters of at least 80% similarity. Clusters that were found in at least 74 of the 75 strains were determined to be core genes, while those found in only some

of the strains were designated as accessory genes. In total, 24% (2448/10200) of the genes were found across all strains (core genome) while 76% (7752/10200) were part of the accessory genome (Figure 2.4). We further classified the core genome by clusters of orthologous groups (COGs) and found that while a large group (21%) had unknown functions, the remaining 79% of the core genome had a widely varied classification spanning 19 other COG categories. Overall the core genome had 33% COGs pertaining to metabolic functions. Particularly interesting was that 8.9% of the core genome was composed of functions in amino acid transport and metabolism (category E), suggesting that this area of metabolism might be particularly conserved over these strains of *A. baumannii*. We also classified the pan genome and note that roughly half could not be COG classified and almost half of that classified portion was classified as having unknown function (Supplementary Figure A.2). This suggests that more robust study and classification of these strains is necessary.

After analyzing the full set of annotated ORFs across the 75 strains, we were particularly interested in applying the iCN718 reconstruction to construct draft strain-specific models of them. To accomplish building these draft models, we determined presence or absence of the 718 genes in the reconstruction of AYE and deleted genes accordingly for the other 74 strains. After this process, we had a measure of the "metabolic pan-genome" as it relates to the genes contained within iCN718. Utilizing the same thresholds, we found that 86% of the genes in iCN718 were considered to be core to all 74 additional strains. Therefore, much of the metabolism represented in iCN718 is maintained in these strains. Three genes were unique to strain AYE within the iCN718 reconstruction: p3ABAYE0029, p2ABYAYE0004, and ABAYE3614. Noting that most of the iCN718 reconstruction was determined to be part of the core metabolic function for all 75 of these strains, we decided to investigate each strain-specific model's metabolic capabilities.

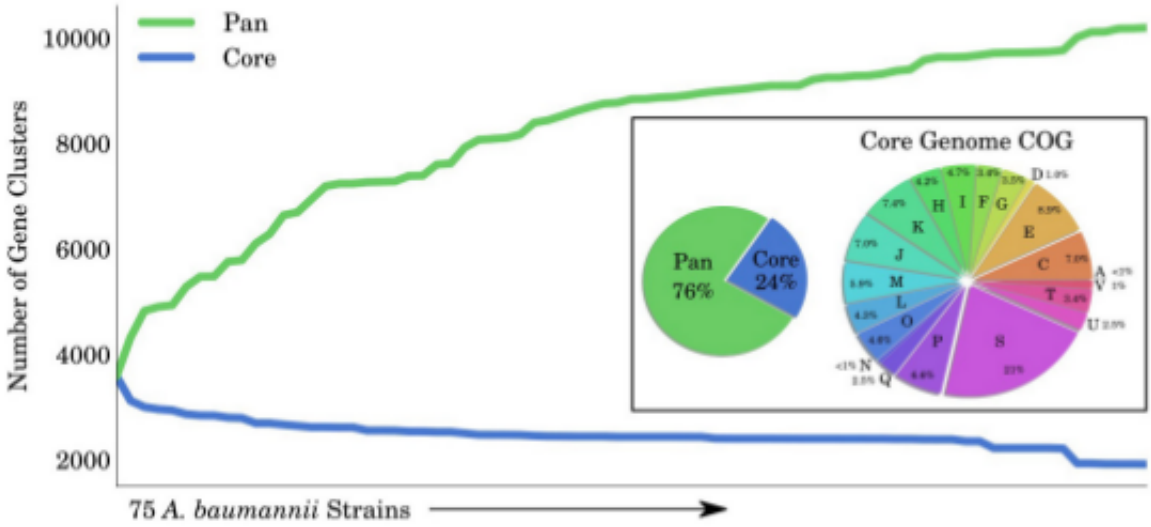

Figure 2.4: Pan and Core Genome of *Acinetobacter baumannii*. The total number of gene clusters in 75 *Acinetobacter baumannii* strains (pan-genome) compared to those that are shared among all strains (core-genome). In total, 76% of the clusters are classified as accessory and 24% as core. The core genome was functionally classified into COG categories. COG categories are as follows: Cellular processes and signaling: D is cell cycle control, cell division, and chromosome partitioning; M is cell wall/membrane/envelope biogenesis; N is cell motility; O is posttranslational modification, protein turnover, and chaperones; T is signal transduction mechanisms; U is intracellular trafficking, secretion, and vesicular transport; V is defense mechanisms; W is extracellular structures; Y is nuclear structure; and Z is cytoskeleton. Information storage and processing: A is RNA processing and modification; B is chromatin structure and dynamics; J is translation, ribosomal structure, and biogenesis; K is transcription; and L is replication, recombination, and repair. Metabolism: C is energy production and conversion; E is amino acid transport and metabolism; F is nucleotide transport and metabolism; G is carbohydrate transport and metabolism; H is coenzyme transport and metabolism; I is lipid transport and metabolism; P is inorganic ion transport and metabolism; and Q is secondary metabolite biosynthesis, transport, and catabolism.

We were additionally interested in analyzing which genes from iCN718 were lost most often (Figure 2.5). The full clustermap of deletions is available in Supplementary Figure A.3. Genes involved in fatty acid metabolism were by far the most highly represented subsystem exceeding the number of genes in the next highest-represented subsystems, butanoate metabolism and folate biosynthesis, by 47 genes.

Originally, only three of the 74 strain-specific models could simulate growth and the

predominantly determining factor of this was the inability to produce lipopolysaccharide (LPS). This result is unsurprising given that LPS is known to vary from strain to strain [18]. The strains that could still synthesize LPS were A1, AB0057, and AB307-0294, suggesting that these strains may have similar LPS compositions to strain AYE. After recognizing LPS as the main limitation to growth for the majority of the strains, we removed LPS from the biomass function for the remaining strains to investigate other properties. With LPS removed, all but four strains could grow. The four strains unable to grow were, as expected, the four strains with the most deletions from the original AYE model. Interestingly, the one strain that was not isolated from a human, SDF, was instead isolated from lice and required 71 more deletions than the next highest dissimilar strain. This suggests that *Acinetobacter* are indeed highly adaptable to varying environments in their metabolic capabilities and that an expanded pan-genome analysis with a higher number of varied strain environments would yield interesting insights.

We then looked at every strain's ability to grow in the same minimal media conditions with sole carbon and nitrogen sources on which iCN718 was originally tested. All of the strains that could grow without LPS in the biomass function maintained the carbon and nitrogen catabolic capabilities exhibited by AYE in iCN718. This analysis is limited in that we are dealing with draft strain specific models, which are all derived from the content common to iCN718. To account for additional capabilities of each strain requires more data and deeper study of these strains. However, this approach demonstrates that with one high-quality reconstruction, insight can be gleaned into a large number of strains from their sequences alone.

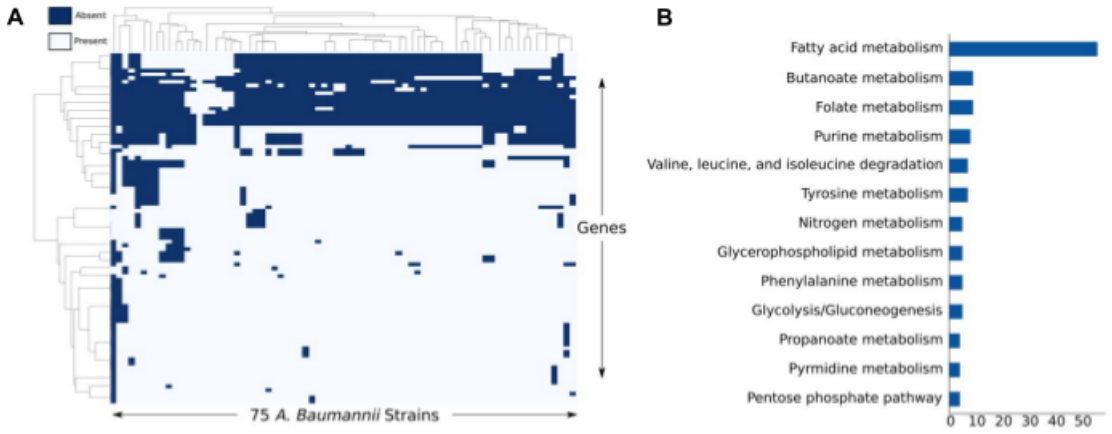

Figure 2.5: Analysis of least conserved genes (A) Clustermap of the genes most deleted from each strain-specific model and (B) the corresponding subsystems of the reactions these genes code for.

2.4 Conclusion

Acinetobacter baumannii is an urgent clinical threat for which treatment is becoming increasingly difficult. High-quality GEMs of strains of *A. baumannii* can be an important tool to accelerate the advancement of new treatments. We updated and improved a previous reconstruction, AbyMBEL891, to produce a new reconstruction, iCN718. We tested iCN718 on multiple gene essentiality datasets as well as phenotypic microarray data. We demonstrated the utility of iCN718 and GEMs to gain further insight into related strains through their sequences alone. iCN718 is in a standardized and curated format that lends itself to further use by the community studying *Acinetobacter*, as well as in future multi-strain reconstructions of diverse *A. baumannii* strains. We demonstrated that iCN718 represents a significant improvement on AbyMBEL891 and a critical step in the progress toward a truly comprehensive knowledge-base for *A. baumannii*. As the knowledge of this organism continues to grow, iCN718 will provide a platform for the integration of further knowledge and data as well as a tool for future investigations.

2.5 Materials and Methods

2.5.1 Reconstructing iCN718

We first obtained a draft metabolic reconstruction of *A. baumannii* AYE utilizing the ModelSeed [11]. AbyMBEL891 was then referenced against this draft reconstruction to compare for the content of each reconstruction. Additional databases (ExPASy, KEGG, MetaNetX, BiGG) were used to refine the reconstruction and obtain a reconstruction utilizing standardized BiGG identifiers [12, 19–21]. The result was a draft reconstruction in BiGG format built upon AbyMBEL891, the draft reconstruction via ModelSeed, and information from the aforementioned databases. To obtain the most accurate final model, this draft reconstruction was then extensively manually curated. This process involved investigating the current literature and rectifying inconsistencies present in the reconstruction. We determined and subsequently filled gaps identified through topological gap analysis and flux-based functional tests. The pathway visualization tool, ESCHER, was instrumental in this gap analysis [22]. We also utilized the GrowMatch algorithm to obtain potential reactions to fill identified gaps [23]. Additionally, the recently published model of *A. baumannii* ATCC 19606, iLP844, was used as an additional resource for cases of conflicting information amongst the aforementioned sources [10]. iLP844 was particularly used to check reaction reversibility. The model content was further improved by comparing it to numerous experimental datasets. In particular, iLP844 was used to confirm reaction reversibility. The model content was further improved by comparing it to numerous experimental datasets and making iterative improvements to increase agreement with experimental data. The manual curation was an iterative process and as such was continuously repeated to yield the highest quality reconstruction possible.

2.5.2 Constraint-Based Modeling

The network reconstruction was converted to a mathematical representation formed from the stoichiometric coefficients of the biochemical reactions. This stoichiometric matrix, S, encapsulates in its columns each mass- and charge-balanced reaction of the network, while each row represents a specific metabolite. The model is assumed to be at homeostatic state (S*V=0). Thermodynamic constraints for network fluxes are incorporated in the form of bounds that incorporate directionality of reactions. The reconstructed model was analyzed with CoBRApy-0.6.1 (COnstraints-Based Reconstruction and Analysis for Python; [24]) and GLPK 4.32 solver. Flux balance analysis (FBA) is a well-established optimization technique and was used in this study. For a primer on FBA, refer to Orth et al. [25].

2.5.3 Gene Essentiality

Gene essentiality predictions were determined by simulating single gene deletions of each applicable gene in the model depending on the dataset in question. Growth of the single gene deletion mutants was predicted using FBA and if, following a gene deletion, there was no growth, this gene was determined to be essential. For all gene-essentiality datasets, the corresponding set of orthologous genes, since no available single gene deletion datasets exist for *A. baumannii* AYE, was obtained via NCBI Bidirectional BLAST (Sayers et al., 2012).

2.5.4 Metabolite Connectivity

The stoichiometric matrices of iCN718 and AbyMBEL891 were used to calculate the metabolite connectivities of every species in each network. The metabolite connectivity is a sum of the number of each reaction a metabolite participates in. Metabolite connectivities were

then ranked from greatest to least connected to form a discrete distribution (Supplementary Figure A.1).

2.5.5 Pan-Genome Analysis

The pan-genome of all 75 completely sequenced strains was constructed by clustering protein sequences based on their sequence homology using the CD-hit package (v4.6). CD-hit clusters protein sequences based on their sequence identity [26]. CD-hit clustering was performed with 0.8 threshold for sequence identity and a word length of 5. A cluster formed by CD-hit is hereon referred to as a gene family. The pan-genome was subdivided into core and accessory genomes. We defined the core genome as gene families that were found in at least 74/75 strains. The subdivided pan-genome was subsequently utilized to identify genes that were part of the core or accessory genome.

Acknowledgements

Chapter 2, in part, is a reprint of material published in: **Norsigian, Charles J.**, Erol Kavvas, Yara Seif, Bernhard O. Palsson, and Jonathan M. Monk. "iCN718, an updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE." *Frontiers in genetics* 9 (2018): 121. The dissertation author was the primary author.

2.6 References

1. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. en. *J. Infect. Dis.* **197,** 1079–1081 (Apr. 2008).

2. Weber, B. S., Harding, C. M. & Feldman, M. F. Pathogenic Acinetobacter: from the Cell Surface to Infinity and Beyond. en. *J. Bacteriol.* **198,** 880–887 (Dec. 2015).

3. Vallenet, D., Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., Bataille, E., Dossat, C., Gas, S., Kreimeyer, A., Lenoble, P., Oztas, S., Poulain, J., Segurens, B., Robert, C., Abergel, C., Claverie, J.-M., Raoult, D., Médigue, C., Weissenbach, J. & Cruveiller, S. Comparative analysis of Acinetobacters: three genomes for three lifestyles. en. *PLoS One* **3,** e1805 (Mar. 2008).

4. Peleg, A. Y., Seifert, H. & Paterson, D. L. Acinetobacter baumannii: emergence of a successful pathogen. en. *Clin. Microbiol. Rev.* **21,** 538–582 (July 2008).

5. Kim, H. U., Kim, T. Y. & Lee, S. Y. Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen Acinetobacter baumannii AYE. en. *Mol. Biosyst.* **6,** 339–348 (Feb. 2010).

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

7. Ebrahim, A., Almaas, E., Bauer, E., Bordbar, A., Burgard, A. P., Chang, R. L., Dräger, A., Famili, I., Feist, A. M., Fleming, R. M. T., Fong, S. S., Hatzimanikatis, V., Herrgård, M. J., Holder, A., Hucka, M., Hyduke, D., Jamshidi, N., Lee, S. Y., Le Novère, N., Lerman, J. A., Lewis, N. E., Ma, D., Mahadevan, R., Maranas, C., Nagarajan, H., Navid, A., Nielsen, J., Nielsen, L. K., Nogales, J., Noronha, A., Pal, C., Palsson, B. O., Papin, J. A., Patil, K. R., Price, N. D., Reed, J. L., Saunders, M., Senger, R. S., Sonnenschein, N., Sun, Y. & Thiele, I. Do genome-scale models need exact solvers or clearer standards? *Mol. Syst. Biol.* **11** (Oct. 2015).

8. Farrugia, D. N., Elbourne, L. D. H., Hassan, K. A., Eijkelkamp, B. A., Tetu, S. G., Brown, M. H., Shah, B. S., Peleg, A. Y., Mabbutt, B. C. & Paulsen, I. T. The complete genome and phenome of a community-acquired Acinetobacter baumannii. en. *PLoS One* **8,** e58628 (Mar. 2013).

9. Gallagher, L. A., Ramage, E., Weiss, E. J., Radey, M., Hayden, H. S., Held, K. G., Huse, H. K., Zurawski, D. V., Brittnacher, M. J. & Manoil, C. Resources for Genetic and Genomic Analysis of Emerging Pathogen Acinetobacter baumannii. en. *J. Bacteriol.* **197,** 2027–2035 (June 2015).

10. Presta, L., Bosi, E., Mansouri, L., Dijkshoorn, L., Fani, R. & Fondi, M. Constraint-based modeling identifies new putative targets to fight colistin-resistant A. baumannii infections. en. *Sci. Rep.* **7,** 3706 (June 2017).

11. Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. & Stevens, R. L. High-throughput generation, optimization and analysis of genome-scale metabolic models. en. *Nat. Biotechnol.* **28,** 977–982 (Sept. 2010).

12. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic Acids Res.* **44,** D515–22 (Jan. 2016).

13. Becker, S. A., Price, N. D. & Palsson, B. Ø. Metabolite coupling in genome-scale metabolic networks. en. *BMC Bioinformatics* **7,** 111 (Mar. 2006).

14. Dorsey, C. W., Tomaras, A. P. & Actis, L. A. Genetic and phenotypic analysis of Acinetobacter baumannii insertion derivatives generated with a transposome system. en. *Appl. Environ. Microbiol.* **68,** 6353–6360 (Dec. 2002).

15. De Berardinis, V., Vallenet, D., Castelli, V., Besnard, M., Pinet, A., Cruaud, C., Samair, S., Lechaplais, C., Gyapay, G., Richez, C., Durot, M., Kreimeyer, A., Le Fèvre, F., Schächter, V., Pezo, V., Döring, V., Scarpelli, C., Médigue, C., Cohen, G. N., Marlière, P., Salanoubat, M. & Weissenbach, J. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. en. *Mol. Syst. Biol.* **4,** 174 (Mar. 2008).

16. 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).

17. 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).

18. Pantophlet, R., Nemec, A., Brade, L., Brade, H. & Dijkshoorn, L. O-antigen diversity among Acinetobacter baumannii strains from the Czech Republic and Northwestern Europe, as determined by lipopolysaccharide-specific monoclonal antibodies. en. *J. Clin. Microbiol.* **39,** 2576–2580 (July 2001).

19. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. en. *Nucleic Acids Res.* **28,** 27–30 (Jan. 2000).

20. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. en. *Nucleic Acids Res.* **31,** 3784–3788 (July 2003).

21. Ganter, M., Bernard, T., Moretti, S., Stelling, J. & Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. en. *Bioinformatics* **29,** 815–816 (Mar. 2013).

22. King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E. & Palsson, B. O. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. en. *PLoS Comput. Biol.* **11,** e1004321 (Aug. 2015).

23. Kumar, V. S. & Maranas, C. D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. en. *PLoS Comput. Biol.* **5,** e1000308 (Mar. 2009).

24. 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).

25. Orth, J. D. & Palsson, B. Ø. Systematizing the generation of missing metabolic knowledge. en. *Biotechnol. Bioeng.* **107,** 403–412 (Oct. 2010).

26. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

Chapter 3

Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase-Producing Multidrug-Resistant *Klebsiella pneumoniae* Clinical Isolates.

3.1 Abstract

The emergence and spread of metallo-beta-lactamase–producing multidrug-resistant *Klebsiella pneumoniae* is a serious public health threat, which is further complicated by the increased prevalence of colistin resistance. The link between antimicrobial resistance acquired by

strains of *Klebsiella* and their unique metabolic capabilities has not been determined. Here, we reconstruct genome-scale metabolic models for 22 *K. pneumoniae* strains with various resistance profiles to different antibiotics, including two strains exhibiting colistin resistance isolated from Cairo, Egypt. We use the models to predict growth capabilities on 265 different sole carbon, nitrogen, sulfur, and phosphorus sources for all 22 strains. Alternate nitrogen source utilization of glutamate, arginine, histidine and ethanolamine among others provided discriminatory power for identifying resistance to amikacin, tetracycline and gentamicin. Thus, genome-scale model based predictions of growth capabilities on alternative substrates may lead to construction of classification trees that are indicative of antibiotic resistance in *Klebsiella* isolates.

3.2 Introduction

The emergence of metallo-beta-lactamase–producing pathogens is a serious challenge to the treatment of clinical infections and a potential public health threat [1]. These pathogens have been identified in the popular news media as "superbugs" because they exhibit multidrug-resistance and can cause infections resistant to all beta-lactams, including last-line options such as carbapenems, as well as most other antibiotics except colistin and sometimes tigecycline [2]. Among multidrug-resistant (MDR) pathogens, six bacterial species have been described as the most threatening, the ESKAPE pathogens [3], which includes *Klebsiella pneumoniae*.

K. pneumoniae is a facultative anaerobic gram-negative bacterium that causes a wide range of clinical diseases including pneumonia, upper respiratory tract infections, wound infections, urinary tract infections and septicemia [4]. Nosocomial infections caused by metallo-β-lactamase–producing *K. pneumoniae* are associated with high rates of morbidity and mortality [5]. This calls for rapid identification of bacteria carrying bla NDM-1 and implementation of

strict infection control measures.

New Delhi metallo-β-lactamase (NDM-1)– producing *K. pneumoniae* have swiftly spread worldwide since an initial report in 2008 [6]. Here, we examined the genomes of four *K. pneumoniae* strains isolated from clinics in Cairo, Egypt. We reconstruct genome-scale models for 2 MDR *Klebsiella pneumoniae* strains (Strains SF and SK), which produce two metallo-β-lactamases (bla NDM-1 and bla VIM-1) and are also colistin resistant. We sequenced these two genomes with two other genomes from strains representing different levels of resistance: one MDR but non-colistin-resistant strain (HM) and a fourth strain (SP) that is not as highly resistant. We then create strain specific genome scale models for each of these four strains as well as an additional 18 publicly available strains to analyze differences in catabolic capabilities in these strains and investigate if these differences can be used to classify resistance phenotypes.

3.3 Results and Discussion

3.3.1 Comparative Genomics of 22 *Klebsiella pneumoniae* Isolates With Defined AMR Phenotypes

We used the PATRIC database [7] to identify complete, single-contig genome sequences that also had experimental evidence of antimicrobial resistance. There were 18 genomes that met this criteria. We supplemented this set with four recently sequenced *K. pneumoniae* strains isolated from patients in Cairo, Egypt collected between 2012 and 2015 [8] (Attia et al. 2019). Three of these isolates are pan-resistant (SF, SK and HM) with two additionally resistant to colistin (SF and SK). A fourth strain, "SP", is multi-drug resistant but sensitive to 10 tested antibiotics. This led to a total set of 22 genomes for comparison (Methods). We assigned

sequence types (ST) to each of the strains using PubMLST [9, 10]. The three colistin resistant strains were found to be part of ST101, known to be a dominant ST for carbapenem resistant *K. pneumoniae* [11]. Next we performed comparative genomics on the full set of 22 strains. We calculated core and pan-genomes for these 22 strains using PanX [12]. The pan-genome consists of all genes found in any of the strains while the core genome consists of genes shared by all strains. The pan-genome for these 22 strains was composed of 10,796 predicted ORFs. Of these, 3,965 are shared amongst all of the strains, forming a core-genome (Figure 3.1). The difference between core and pan-genomes is called the accessory genome and consists of genes that make the individual strains unique. In this case there are 4,026 accessory genes and 2,805 unique genes (those found in only 1 strain). We compared the presence of different accessory genes across the strains. Hierarchical clustering of the accessory gene contents demonstrated stratification by sequence type (Supplementary Figure B.1). The three pan-resistant strains from ST101 (SF, SK and HM) clustered together based on accessory gene content. Next we used the CARD database [13]genome for AMR encoding processes to form a "resistome" of the strains. In total there were 122 predicted AMR encoding genes or mutations across all 22 strains, with 12 shared by all strains, 72 variably present across the strains and 35 unique to single strains (Figure 3.1). We found that hierarchical clustering of AMR determinants also grouped strains by sequence type (Figure 3.1). Next we performed an in-depth analysis of the ST101 strains.

3.3.2 Focused Genomic-Analysis of Four *Klebsiella pneumoniae* Isolates From Cairo, Egypt

Genomic analyses were performed to determine the genetic similarity of the four *K. pneumoniae* isolates from Cairo, with the model *K. pneumoniae* strain MGH78578 included as

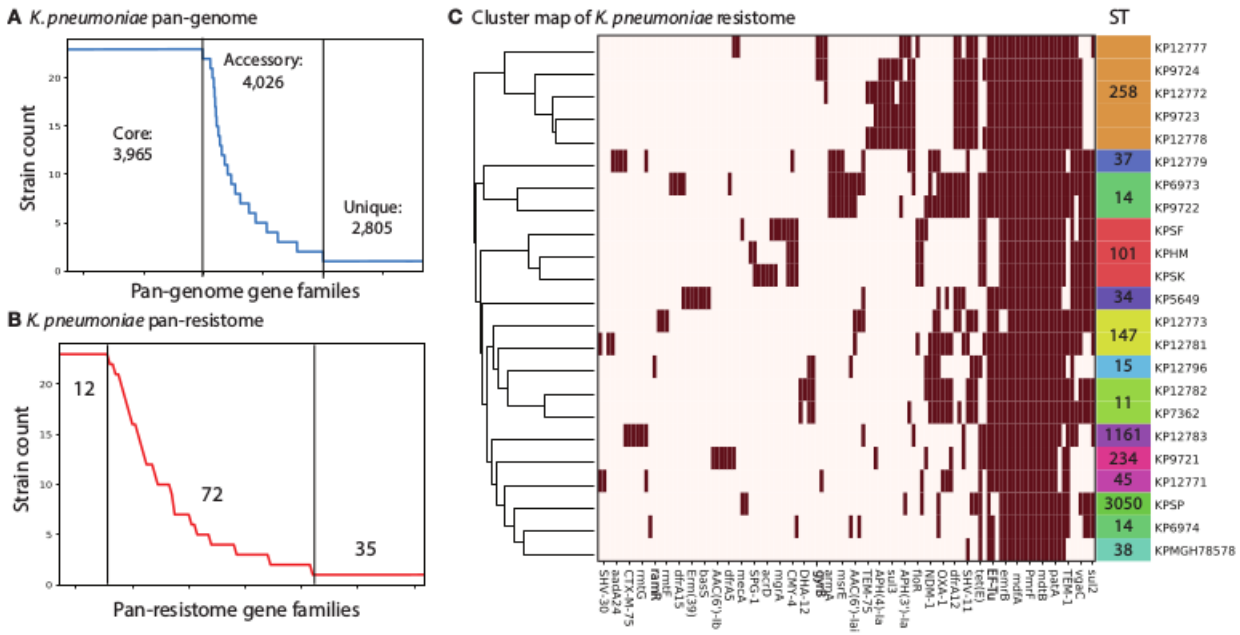

Figure 3.1: Comparative genomics analysis of 22 *K. pneumoniae* strains. (A) The calculated pan-genome of the 22 KP strains consisted of 10,999 gene families, 3,965 of which were core, 4,026 were accessory and 2,805 unique. The full clustermap of the accessory genome is available as Supplementary Figure B.1. (B) A pan "resistome" was constructed by mapping the genomes to the CARD database. A total of 12 AMR determinants were shared by all strains, 72 were strain specific and 35 were unique to individual strains. (C) Hierarchical clustering of the contents of the pan-resistome shows that KP strains group into sequence types (ST) based solely on resistance encoding mechanisms. The gyrB variants were recorded as resistant genes based on 100% identity to resistant-conferring mutants in CARD.

a reference (Genbank ID: CP000647.1). A core-genome phylogenetic tree was constructed using PanX [12] and demonstrated that the three pan-resistant strains were most similar to each other (Figure 3.2). We included the model-strain, *K. pneumoniae* MGH78578 as a reference and this strain was most dissimilar compared to the other resistant strains. Next, we constructed core and pan-genomes for these five strains. The pan-genome consists of all genes found in any of the five strains while the core genome consists of genes shared by all five strains. The pan genome was composed of 6,879 predicted ORFs across all five strains with 4,336 shared amongst all of the strains, forming a core-genome and 2,549 accessory genes. We compared the presence of

different accessory genes across the strains. Hierarchical clustering of the accessory gene contents agreed with the whole-genome phylogeny to show that the three pan-resistant strains (SF, SK and HM) are most similar while SP and MGH are more dissimilar in terms of shared accessory genes (Figure 3.2).

We hypothesized that genetic background and gene portfolio of individual strains may have a role in acquisition and spread of antibiotic resistance. Thus, we identified the shared and strain-specific genes amongst these five strains. In total, 4336 genes were shared amongst all five *K. pneumoniae* strains with 536 genes unique to strain SP, 340 genes unique to MGH and 1330 genes unique to the three pan-resistant strains (Figure 3.2). In total the three pan-resistant strains shared 5,411 genes with each other while another 541 were uniquely present across these three strains (Figure 3.2). More than one-third of the uniquely present genes (35%) were predicted to have metabolic functions, potentially indicating that nutrient niche and unique metabolic capabilities may influence acquisition of antimicrobial resistance determinants. Genome-scale models of metabolism have demonstrated utility at systematically categorizing the metabolic capabilities of strains in a species [14–16]. To further investigate this hypothesis we set out to construct genome-scale models of the five strains as well as other publically-available strains with antimicrobial profiling data.

3.3.3 Diverse Catabolic Capabilities of Multiple *Klebsiella pneumoniae* Strains

We used the experimentally validated genome-scale metabolic reconstruction, iYL1228 [17], as a platform to investigate the metabolic differences amongst our group of isolates. iYL12228 is a reconstruction for *K. pneumoniae* MGH78578 and provided a valuable resource to

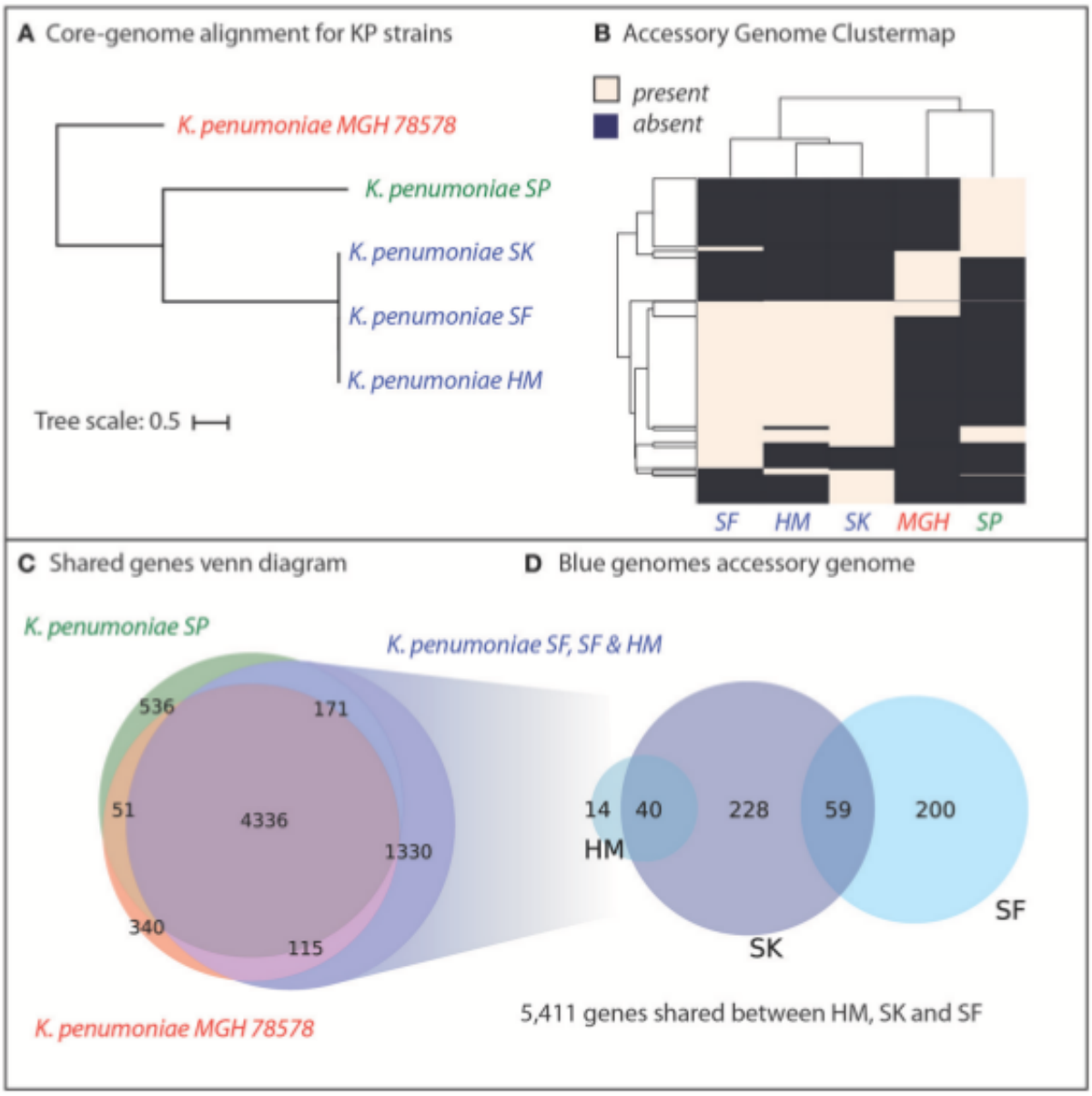

Figure 3.2: Genomic analyses of five *K. pneumoniae* isolates including four isolated in Cairo, Egypt (SP, SK, SF, HM) (A) Core-genome phylogenetic tree demonstrates that three of the *K. pneumoniae* isolates (SK, SF, HM) are most similar to each other, with SP and the model strain MGH78578 more distantly related. (B) Hierarchical clustering of the accessory genome of the five strains demonstrates that the three closely related KP strains also share the most genes. (C) The core genome of all five strains is composed of 4,366 genes shared by all five strains. (D) The three strains, SD, SK, and HM possess an additional 1,330 unique genes not shared by SP or MGH78579.

link the genetic information of other strains to defined metabolic reactions (Methods). We first built draft models of all strains using sequence similarity. Following that we added additional metabolic content identified through the use of DETECT v2 [18], an enzyme annotation tool. This process allowed us to include additional metabolic processes unique to each of the strains. Of these strains, initially 10 of the draft models could not solve for biomass. We used GrowMatch [19] to gapfill these networks and found that the removal of the reactions TDPDRE encoded for by gene KPN_02494 or KPN_02488 and TDPDRR encoded for by gene KPN_02495 or KPN_02489 was the cause. These reactions are directly involved in the production of DTDP-L-rhamnose, a metabolite directly required for biomass production in iYL12228. We hypothesized that either keeping these reactions in the network or removing DTDP-L-rhamnose from the biomass function would restore growth of these models. Given that the homologous genes from strain MGH78578 were not present in the other strains, we opted to remove DTDP-L-rhamnose from the biomass function for the models of these strains. This assumption is valid given that DTDP-L-rhamnose is involved in the biosynthesis of peptidoglycan and it is likely that these strains have variant peptidoglycan composition [20, 21]. Additionally, through the gapfilling process we identified that one strain, KP9721, was predicted to be auxotrophic for proline and as such in the following analyses this model was supplemented with proline in the in silico media. Using our 22 total models derived from iYL12228 we sought to analyze the various catabolic capabilities present across the strains. It is worth noting that these catabolic capabilities are predictive and could be used in conjunction with future study of actual phenotypes. The quality of the models could be improved in the future by validating with experimental data such as gene essentiality or phenotypic arrays such as Biolog should that data become available.

To interrogate each of the strain's catabolic capabilities we simulated for biomass produc-

tion in minimal media conditions (in silico M9 media) and alternated carbon, nitrogen, sulfur, and phosphorus sources to simulate each strain's ability to grow on a variety of compounds (Figure 3.3). The simulations for carbon, nitrogen, and sulfur provided some interesting differences strain to strain whereas capabilities for various phosphorus sources were largely conserved across the entire group (Supplementary Figures B.2 and B.3). For carbon sources one apparent difference is that the KP9721 and KP12783 models lack the ability to use maltose and any of its derivatives (maltotriose, maltotetrose, maltohexaose, maltopentaose) whereas all the other models can utilize these sugars. These models also are the only two unable to catabolize glutamate as a carbon source. Further, KP12783 uniquely cannot utilize ascorbate or lyxose. Another model with unique loss of capabilities relative to the others was KP5649 being unable to grow on fucose, rhamnose, and glucarate and the only two strains unable to use glucosamine or mannose were KP12781 and KP12773. The following compounds are unable to be used by various small groups of strains: ribose, mannitol, glyceraldehyde, glutamine, D-Alanyl-D-alanine, and galacturonate. Conversely, the following compounds can be used by only various smaller groupings of the strains: glycine, prolinylglycine, and 2-Dehydro-3-deoxy-D-gluconate.

The model-predicted growth capabilities on nitrogen sources were slightly less varied than for carbon. Given the predicted auxotrophy for proline in KP9721, we omit its inability to use the majority of other nitrogen sources in the following summary. Both models for KP5649 and KP12779 fail to utilize a large number of nitrogen sources (Figure 3.3). Models of KP9723, 12777, KP9722, KP12778, and KP12772 all could not make use of prolinylglycine, D-Alanyl-D-alanine, or cys-glycine. KP9723 was additionally the only strain unable to use arginine or agmatine. Glucosamine could not be used by KP12781 or KP12773. Histine could not be used by KP127777. Finally, ethanolamine could not be used by KP6974 or KP12771. Ability to utilize

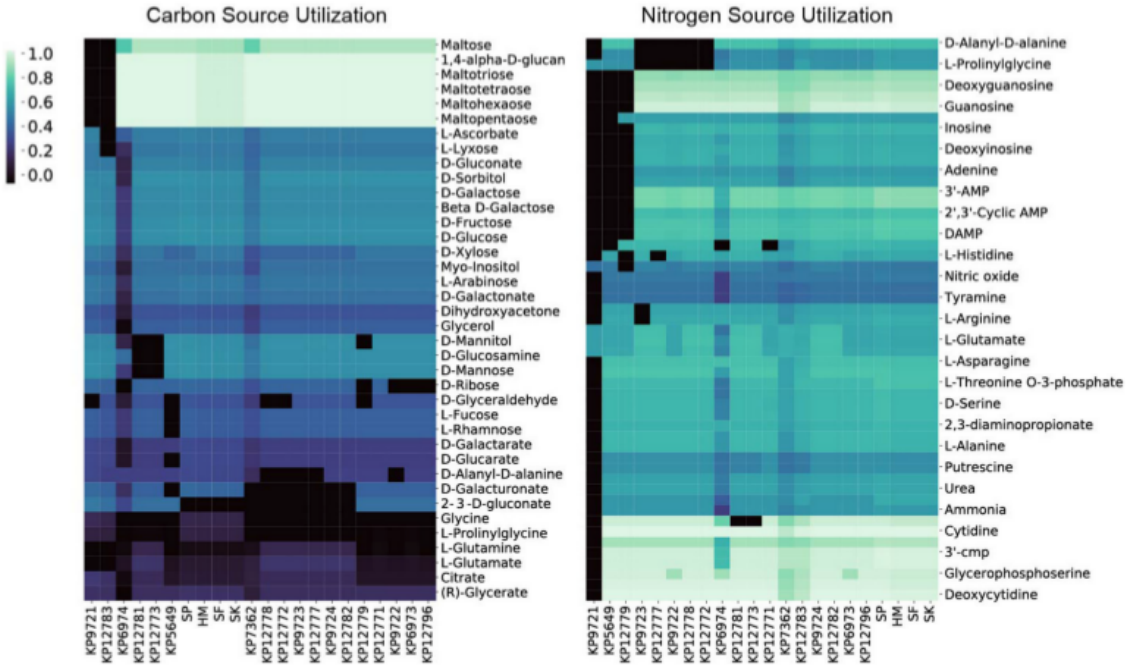

Figure 3.3: The 22 *in silico* models predicted relative carbon and nitrogen source utilization. By simulating in minimal media and swapping only the carbon or nitrogen source the predicted catabolic capabilities were calculated. The resulting *in silico* predicted biomass objective flux for each strain on the various sources is reported and hierarchically clustered here. Interestingly, in both the case of carbon and nitrogen source utilization the four isolates from Egypt (SP, SF, SK, HM) all cluster together.

alternate nitrogen sources is interesting in light of the fact that elevated blood urea nitrogen levels are a biomarker of *K. pneumoniae* pathology and associated with a poor prognosis [22, 23]. Also, *Klebsiella* is the only genus in the family enterobacteriaceae able to fix nitrogen in the atmosphere and convert it to ammonia and amino acids using an energy intensive nitrogenase [24, 25], further highlighting the importance of this element in *Klebsiella* lifestyle and niche.

Lastly, there were far fewer sulfur sources available to test than carbon or nitrogen but this analysis still provided some interesting differences amongst the strains. Chiefly, only the models of strains isolated from Egypt (SP, HM, SF, and SK) could utilize ethanesulfonate, isethionic acid, or sulfoacetate as sulfur sources. Interestingly, only SP was predicted to be capable of using

methionine as a sulfur source whereas models for SP, KP12777, KP9723, KP12778, KP12772, KP9724, KP12781, KP12783, KP12796 and KP12771 could all use Methyl-L-methionine. Lastly KP9722, KP12777, KP9723, KP12778, KP12772 were all predicted to be unable to utilize glutathione and cys-glycine.

3.3.4 Substrate Usage to Classify Antimicrobial Resistance Phenotypes

After using the draft models to generate predicted catabolic capabilities for all 22 strains we sought to see if these catabolic capabilities were correlated with the antimicrobial resistance phenotypes of the strains. As previously noted, strains SF and SK are both MDR as well as colistin resistant, HM is MDR but not colistin resistant, and SP is susceptible to a number of drugs. The 18 strains we included from PATRIC were selected partly on the availability of experimental AMR profiling. We used this data from PATRIC and the results of both disk diffusion and agar dilution methods on our four clinical isolates (Table 3.1, Supplementary Tables B.1 and B.2) to construct the resistance profiles for which drug data existed for all the strains (Supplementary Figure B.4). Unfortunately, the strains from PATRIC do not have conclusive profiling of colistin resistance. It was immediately apparent that 7 of the strains were resistant to all 16 drugs. Additionally, 7 of the drugs were resisted by all 22 strains. Of the remaining drugs tetracycline, amikacin, and gentamicin had the most strains either susceptible or intermediately resistant. As such these drugs were considered for further analysis. Interestingly, all three of these drugs target protein synthesis and both amikacin and gentamicin are both aminoglycosides [13]. Yet, the group of strains had varied resistance phenotypes to these same class drugs.

Table 3.1: Antimicrobial resistance profile of the isolated *K. pneumoniae* strains determined by disk diffusion.

Antibiotic	SF	SK	HM	SP
Amikacin	R	R	R	R
Amoxicillin/ clavulanic acid	R	R	R	R
Ampiciliin	R	R	R	R
Aztreonam	R	R	R	R
Cefaclor	R	R	R	R
Cefepime	R	R	R	S
Cefotaxime	R	R	R	R
Cefoxitin	R	R	R	S
Ceftazidme	R	R	R	R
Ceftriaxone	R	R	R	R
Cefuroxime sodium	R	R	R	R
Chloramphenicol	R	R	R	R
Colistin	R	R	S	S
Ertapenem	R	R	R	S
Gentamicin	R	R	R	S
Imipenem	R	R	R	R
Lomefloxacin	R	R	R	S
Meropenem	R	R	R	S
Netlimicin	R	R	R	R
Nitrofurantoin	R	R	R	S
Piperacillin	R	R	R	S
Piperacillin/Tazobactam	R	R	R	R
Trimethoprim/Sulfamethoxazole	R	R	R	R
Tetracycline	R	R	R	S

To determine whether model-predicted metabolic capabilities could be linked to antibiotic resistance, we constructed classification trees using scikit-learn [26] for tetracycline, amikacin, and gentamicin resistance based on the relative in silico predicted biomass yields on various carbon or nitrogen sources (Supplementary Figures B.5 - B.9). We limited our analyses to carbon and nitrogen sources because the number of model-predictions for these compounds greatly exceeds those for sulfur sources. Based on simulated growth phenotypes, we sought to determine whether model-predicted growth capabilities could stratify strains that were resistant, intermediate, or susceptible to a given drug. Interestingly, the trees based on nitrogen sources were able to classify

the strains at lower tree depths than other nutrient sources (Figure 3.4). In particular the trees for tetracycline and amikacin both possessed the same right branching architecture based on variant usage of arginine and histidine as nitrogen sources. In both cases 6 strains that are then classified by their usage of these two amino acids are KP9724, KP12778, KP9723, KP12781, and KP12777 and in the case of tetracycline also SP and KP127771.

Given this shared grouping of strains across the different drug phenotype profiles we looked back to the draft models to see which genes that were lost could be attributed to this grouping. Interestingly, genes with homology to KPN_00956 and KPN_00282 were both deleted from all of these strains and either no other strains or only 3 other strains in the case of KPN_00282. Both of these genes participate in the gene product rules for over 200 transport reactions in the reconstruction, but these reactions have other genes maintained within the gene product rule as well. Lastly, it is interesting to note that the complete inability to use arginine by the model for strain KP9723 as well as the complete inability to use histidine by the model for KP12777 in both classification schema are critical for separating the AMR phenotypes. One limitation of this methodology is the small sample size [27] as well as the use of relative biomass yield for the growth phenotypes. This leads to some of the classification trees being overly deep or making branches at very small differences in biomass flux. Further extensive studies of *Klebsiella pneumoniae* with increased diversity of strains in capabilities as well as drug resistance could provide future valuable delineating features. Nevertheless these initial results are promising and demonstrate that it could be possible to construct a robust classification schema of AMR capabilities based on model predicted growth capabilities in the future.

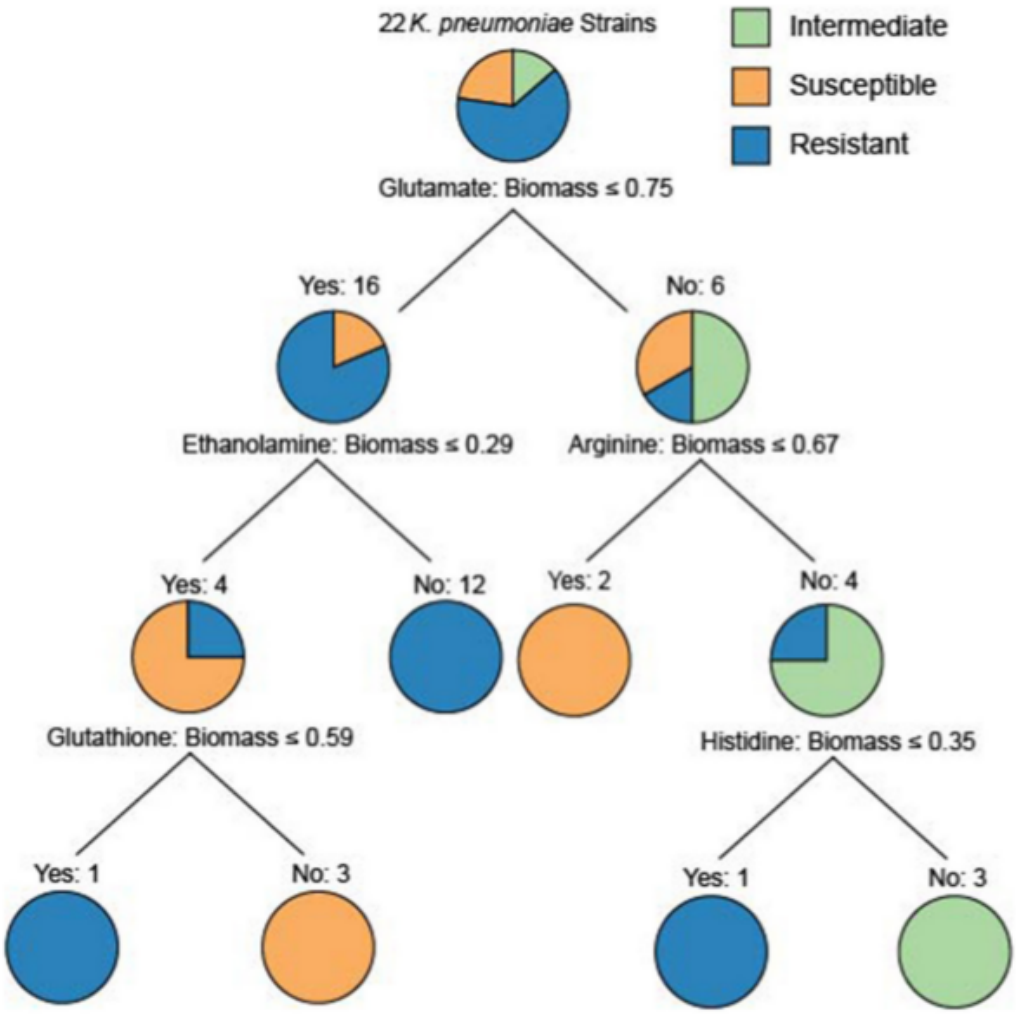

Figure 3.4: Classification tree built based upon nitrogen source utilization classifying the amikacin resistance phenotypes. Interestingly, the ability to utilize glutamate initially discriminates the majority of the resistant strains. The right branching tree architecture utilizes arginine and histidine utilization to quickly discriminate the proper groupings of intermediate, susceptible, and resistant strains. These trees are an effort to examine the ability of differential predicted catabolic capabilities to discriminate varying resistance phenotypes of the strains. For trees generated using carbon source utilization as well as for the resistance phenotypes for tetracycline and gentamicin see Supplementary Figures B.5 - B.9

3.4 Conclusion

K. pneumoniae continues to be a serious threat and increasing antimicrobial resistance is exacerbating this problem [28]. We used genome scale metabolic models to demonstrate

that there exist differences in predicted catabolic capabilities amongst a group of MDR strains. Through this systems biology approach we also demonstrated the possibility of constructing a classification schema for antimicrobial resistance based on these capabilities. The robustness of this strategy could be improved by increasing the number of strains with the pertinent resistance phenotype data included. GEMs could be used in the future to delineate which metabolic capabilities are potential drivers of infection niches for *K. pneumoniae*.

3.5 Materials and Methods

3.5.1 Construction of Draft Strain-Specific Models

The sequences of the 22 selected strains were all downloaded from PATRIC and re-annotated using PROKKA v.1.2 [29], They were then compared based on annotated ORF amino acid sequence similarity using NCBI bidirectional BLAST. A 0.9 threshold was used for assigning orthologs. Genes with a score below 90 were deleted from the strain-specific model. In this manner derivative draft strain-specific models of all 22 strains were generated with the designated orthologous genes removed from the base model iYL12228. All 22 strain-specific models are available as json files. Gene names within the model are as per the locus tags in the original base model in the 18 strains acquired from the PATRIC database. The models for SP, SF, SK, HM had additional content curated through the use of the DETECT v2 algorithm and gene names are as per each strain's locus tags. Further the change of DTDP-rhamnose in the biomass equation is as described in the main text amongst the strains and this is the only change in biomass equation amongst the strains.

3.5.2 *In silico* Growth Simulations

For the in silico growth simulations, the following minimal media similar to M9 minimal media was used: glucose, calcium, chloride, carbon dioxide, cobalt, copper, iron, hydrogen, magnesium, manganese, molybdate, sodium, oxygen, ammonia, phosphate, zinc, tungstate, and sulfate. The in silico media used with corresponding exchange reactions and lower bounds is available as Supplementary Table 3. From this minimal media the following metabolites glucose, ammonia, phosphate, and sulfate were removed to evaluate other sources of carbon, nitrogen, phosphorus, and sulfur respectively. This analysis involves removing each of these compounds from the media (setting lower bound to zero) and testing other compounds using flux balance analysis to determine if these compounds can support growth. In the case of strain KP9721, which was predicted to be auxotrophic for proline, the media was supplemented with proline. Growth versus no growth determinations in all conditions were determined through flux balance analysis on each described nutrient condition, optimizing for the biomass function. Biomass objective flux of greater than zero designated a metabolites capable of growth supporting. For further information and tutorials on these methods see the COBRApy documentation (`https://cobrapy.readthedocs.io/en/latest/`).

3.5.3 Construction of Classification Trees

Before building the trees we filtered the carbon and nitrogen sources to exclude the compounds that were overly similar in in silico biomass yield across all 22 strains based on standard deviation of the biomass objective flux across the 22 strains for a given source. Classification trees were calculated using relative biomass objective flux found through flux balance analysis for each strain on the tested nutrient sources. These catabolic capabilities were used to classify the

strains into their resistance phenotypes: resistant, intermediate, or susceptible (Supplementary Figure B.4) for a given single drug. The decision tree classifier from sklearn was used to generate the trees with no binarization.

3.5.4 Nucleotide Sequence Accession Numbers

The four isolates that were sequenced and their annotations are deposited in NCBI as RXLW00000000, RXLX00000000, RXLY00000000 and RXLV00000000 as well as in the PATRIC database (http://www.patricbrc.org) under the following genome IDs: 573.18994, 573. 19098, 573. 18993, 573. 18996 for SF, HM, SK, SP, respectively. Additionally, the 18 previously publicly available stains were downloaded from PATRIC and used in this study have the accession numbers: 573.12771, 573.12772, 573.12773, 573.12777, 573.12778, 573.12779, 573.12781, 573.12782, 573.12783, 573.12796, 573.5649, 573.6973, 573.6974 ,573.7362, 573.9721, 573.9722,573.9723, 573.9724.

3.5.5 Resistance Profiling of 4 Clinical Isolates From Cairo, Egypt

Antimicrobial resistance profiles (Table 3.1) were determined by the Kirby Bauer disk diffusion method [30], and their minimum inhibitory concentrations (MICs) were determined by the agar dilution method to confirm their resistance profile.

3.5.6 Identification of AMR encoding genes

The CARD RGI tool [13] version 3.2.0 with database version 1.1.8 was used to identify genetic determinants of antimicrobial resistance. All identified determinants are available as Supplementary Figure B.10.

3.5.7 MIC Screens

Determination of MIC was performed according to CLSI guidelines described in [31] and [32]sing sterile U shaped 96 well microtiter plates. Each antibiotic was prepared by diluting the powder in water for injection (WFI) as the solvent and the diluent. All antibiotics were purchased from Sigma except Ertapenem, purchased as an Invanz vial from Merck Co. USA. The powder of the drug equivalent to 26.1 mg in case of ertapenem, 3.26 mg in case of meropenem and colistin and 105.2 mg in case of ceftazidime and cefotaxime was dissolved in 20 ml WFI forming a stock solution (solution A) of concentration 1280 µg/ml for ertapenem ,160 µg/ml meropenem and colistin and 5120 µg/ml for ceftazidime and cefotaxime respectively. Solution (B) of concentrations 128, 16 and 512 µg/ml was prepared by diluting 1ml of each solution (A) with 9 ml WFI. Preparation of the 2 fold dilutions A series of 2-fold dilutions was prepared as recommended by (Amsterdam, 2005) by using solution (B) from each stock solution. Inoculum was prepared by selecting several discrete colonies, usually three to five, subcultured in the inoculum growth broth, to avoid single colony variance. The inoculum was cultured in Mueller Hinton broth (MHB), the same broth medium used for the test, incubated at 37°C for 2-6 hours until turbidity is equal or exceed the turbidity of 0.5 McFarland, then the optical density of the bacterial suspension was adjusted using spectrophotometer at a wavelength of 625 nm to the O.D of 0.08-0.13 which approximates a 0.5 McFarland standard. The adjusted culture was then diluted 1:100 times with Muller-Hinton broth (MHB), to bring the inoculum density to the range of 105 to 106 CFU/ml. A set of the 11 prepared antibiotic dilutions for each antibiotic were allowed to warm at room temperature prior to use. The wells of the 96 well microtiter plate were filled with 50 µl from each dilution. The column number 12 was filled with 100 µl MHB for the growth control for each isolate. Each well in the same row was filled with 50 µl

of the tested inoculum. For each experiment, an additional row was left for negative control by adding 100 µl of MHB to the different antibiotic dilutions. The plates were covered with lid. Incubation of the microtiter plate at 37°C for 16-20 hours. Microdilution trays were prepared each day they were used and Unused thawed dilutions were discarded and never refrozen. The plates were read visually on a dark background. The endpoint MIC was the lowest concentration of drug at which the tested microorganism did not show a visible growth. The MIC values of each tested antibiotic against the selected *Klebsiella pneumoniae* isolates are listed in Supplementary Tables B.1 and B.2. The other reported antibiotics were measured using disc diffusion and thus do not have a reported MIC. Instead we only report resistance and susceptibility based on the manufacturers instructions.

Acknowledgements

Chapter 3, in part, is a reprint of material published in: **Norsigian, Charles Joseph**, Heba Attia, Richard Szubin, Aymin Yassin, Bernhard O. Palsson, Ramy K. Aziz, and Jonathan Monk. "Comparative Genome-scale Metabolic Modelling of Metallo-beta-Lactamase–producing Multidrug-resistant Klebsiella pneumoniae Clinical Isolates." *Frontiers in cellular and infection microbiology* 9 (2019): 161. The dissertation author was the primary author.

3.6 References

1. Pitout, J. D. D. & Laupland, K. B. Extended-spectrum β-lactamase-producing Enterobacteriaceae: an emerging public-health concern. *Lancet Infect. Dis.* **8,** 159–166 (Mar. 2008).

2. Kumarasamy, K. K., Toleman, M. A., Walsh, T. R., Bagaria, J., Butt, F., Balakrishnan, R., Chaudhary, U., Doumith, M., Giske, C. G., Irfan, S., Krishnan, P., Kumar, A. V., Maharjan, S., Mushtaq, S., Noorie, T., Paterson, D. L., Pearson, A., Perry, C., Pike, R., Rao, B., Ray, U., Sarma, J. B., Sharma, M., Sheridan, E., Thirunarayan, M. A., Turton, J., Upadhyay, S., Warner, M., Welfare, W., Livermore, D. M. & Woodford, N. Emergence

of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. en. *Lancet Infect. Dis.* **10,** 597–602 (Sept. 2010).

3. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. en. *J. Infect. Dis.* **197,** 1079–1081 (Apr. 2008).

4. Broberg, C. A., Palacios, M. & Miller, V. L. Klebsiella: a long way to go towards understanding this enigmatic jet-setter. en. *F1000Prime Rep.* **6,** 64 (Aug. 2014).

5. Pitout, J. D. D., Nordmann, P. & Poirel, L. Carbapenemase-Producing Klebsiella pneumoniae, a Key Pathogen Set for Global Nosocomial Dominance. en. *Antimicrob. Agents Chemother.* **59,** 5873–5884 (Oct. 2015).

6. Bushnell, G., Mitrani-Gold, F. & Mundy, L. M. Emergence of New Delhi metallo-β-lactamase type 1-producing enterobacteriaceae and non-enterobacteriaceae: global case detection and bacterial surveillance. en. *Int. J. Infect. Dis.* **17,** e325–33 (May 2013).

7. 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).

8. Attia, H., Szubin, R., Yassin, A. S., Monk, J. M. & Aziz, R. K. Draft Genome Sequences of Four Metallo-Beta-Lactamase-Producing Multidrug-Resistant Klebsiella pneumoniae Clinical Isolates, Including Two Colistin-Resistant Strains, from Cairo, Egypt. *Microbiol Resour Announc* **8,** e01418–18 (2019).

9. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. en. *BMC Bioinformatics* **11,** 595 (Dec. 2010).

10. Seemann, T. *mlst*

11. Mammina, C., Bonura, C., Aleo, A., Fasciana, T., Brunelli, T., Pesavento, G., Degl'Innocenti, R. & Nastasi, A. Sequence type 101 (ST101) as the predominant carbapenem-non-susceptible Klebsiella pneumoniae clone in an acute general hospital in Italy. en. *Int. J. Antimicrob. Agents* **39,** 543–545 (June 2012).

12. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. en. *Nucleic Acids Res.* **46,** e5 (Jan. 2018).

13. Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., Johnson, T. A., Brinkman, F. S. L., Wright, G. D. & McArthur, A. G. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. en. *Nucleic Acids Res.* **45,** D566–D573 (Jan. 2017).

14. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

15. 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. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

16. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

17. Liao, Y.-C., Huang, T.-W., Chen, F.-C., Charusanti, P., Hong, J. S. J., Chang, H.-Y., Tsai, S.-F., Palsson, B. O. & Hsiung, C. A. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J. Bacteriol.* (2011).

18. Nursimulu, N., Xu, L. L., Wasmuth, J. D., Krukov, I. & Parkinson, J. Improved enzyme annotation with EC-specific cutoffs using DETECT v2. en. *Bioinformatics* (May 2018).

19. Kumar, V. S. & Maranas, C. D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. en. *PLoS Comput. Biol.* **5,** e1000308 (Mar. 2009).

20. Pan, Y.-J., Lin, T.-L., Chen, C.-T., Chen, Y.-Y., Hsieh, P.-F., Hsu, C.-R., Wu, M.-C. & Wang, J.-T. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp. en. *Sci. Rep.* **5,** 15573 (Oct. 2015).

21. Shu, H.-Y., Fung, C.-P., Liu, Y.-M., Wu, K.-M., Chen, Y.-T., Li, L.-H., Liu, T.-T., Kirby, R. & Tsai, S.-F. Genetic diversity of capsular polysaccharide biosynthesis in Klebsiella pneumoniae clinical isolates. en. *Microbiology* **155,** 4170–4183 (Dec. 2009).

22. Yasin, F., Assad, S., Talpur, A. S., Zahid, M. & Malik, S. A. Combination Therapy for Multidrug-Resistant Klebsiella Pneumoniae Urinary Tract Infection. en. *Cureus* **9,** e1503 (July 2017).

23. Chang, S. W., Yen, D. H., Fung, C. P., Liu, C. Y., Chen, K. K., Tiu, C. M., Wang, L. M. & Lee, C. H. Klebsiella pneumoniae renal abscess. en. *Zhonghua Yi Xue Za Zhi* **63,** 721–728 (Oct. 2000).

24. Leisy Azar, S. & Ebadi, A. R. Examining the Pattern of Susceptibility and Antibiotic Resistance in Klebsiella pneumoniae Strains Isolated from Urine Samples of Children with Urinary Tract Infections from the Children's Hospital of Tabriz in 2015. *Br Biomed Bull* **05** (2017).

25. Dixon, R., Kennedy, C., Kondorosi, A., Krishnapillai, V. & Merrick, M. *Complementation analysis of Klebsiella pneumoniae mutants defective in nitrogen fixation* 1977.

26. 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).

27. Chicco, D. Ten quick tips for machine learning in computational biology. en. *BioData Min.* **10,** 35 (Dec. 2017).

28. Kaplan, S. L. Review of antibiotic resistance, antibiotic treatment and prevention of pneumococcal pneumonia. *Paediatr. Respir. Rev.* **5,** S153–S158 (Jan. 2004).

29. Seemann, T. Prokka: rapid prokaryotic genome annotation. en. *Bioinformatics* **30,** 2068–2069 (July 2014).

30. Bauer, A. W., Kirby, W. M., Sherris, J. C. & Turck, M. Antibiotic susceptibility testing by a standardized single disk method. en. *Am. J. Clin. Pathol.* **45,** 493–496 (Apr. 1966).

31. Amsterdam, D. *Susceptibility testing in liquid media in antibiotics in laboratory medicine, Victor Lorian (ed) Lippincott, Williams and Wilkins* tech. rep. (ISBN 0-7817-4983-2, 2005).

32. Clsi, C. Performance standards for antimicrobial susceptibility testing; twenty-fourth informational supplement. *M100-S24 January* (2014).

Chapter 4

Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence

4.1 Abstract

Hospital acquired *Clostridioides (Clostridium) difficile* infection is exacerbated by the continued evolution of *C. difficile* strains, a phenomenon studied by multiple laboratories using stock cultures specific to each laboratory. Intralaboratory evolution of strains contributes to

interlaboratory variation in experimental results adding to the challenges of scientific rigor and reproducibility. To explore how microevolution of *C. difficile* within laboratories influences the metabolic capacity of an organism, three different laboratory stock isolates of the *C. difficile* 630 reference strain were whole genome sequenced and profiled in over 180 nutrient environments using phenotypic microarrays. The results identified differences in growth dynamics for 32 carbon sources including trehalose, fructose and mannose. An updated genome-scale model for *C. difficile* 630 was constructed and used to contextualize the 28 unique mutations observed between the stock cultures. The integration of phenotypic screens with model predictions identified pathways enabling catabolism of ethanolamine, salicin, arbutin, and N-acetyl-galactosamine that differentiated individual *C. difficile* 630 laboratory isolates. The reconstruction was used as a framework to analyze the core-genome of 415 publicly available *C. difficile* genomes and identify areas of metabolism prone to evolution within the species. Genes encoding enzymes and transporters involved in starch metabolism and iron acquisition were more variable while *C. difficile* distinct metabolic functions like Stickland fermentation were more consistent. A substitution in the trehalose PTS system was identified with potential implications in strain virulence. Thus, pairing genome-scale models with large-scale physiological and genomic data enables a mechanistic framework for studying the evolution of pathogens within microenvironments and will lead to predictive modeling to combat pathogen emergence.

4.2 Introduction

Clostridioides (Clostridium) difficile continues to be a leading cause of hospital-borne infection, adversely affecting patient health as well as causing significant healthcare costs [1]. The continued evolution of *C. difficile* strains to both antibiotic resistance and survival in the host

greatly increases the challenges of treatment [2]. *C. difficile* infection (CDI) occurs following the disruption of the host microbiota after treatment with antibiotics and instances of subsequent recurrent infections are common, often presenting with more severe symptoms [3]. In the absence of the natural microbiota, opportunistic, toxigenic strains of *C. difficile* flourish and produce enterotoxins resulting in the observed patient symptoms. These symptoms are wide-ranging and vary from completely asymptomatic to antibiotic-associated diarrhea to pseudomembranous colitis and even death. Frighteningly, the rate of success for commonly used antibiotics metronidazole and vancomycin is steadily falling [4].

Studying this deadly pathogen in the laboratory requires well characterized stock strains. Unfortunately, the evolution of stock cultures used in laboratory experiments has recently emerged as a major concern. This evolution can lead to the accumulation of genetic changes that have relevant physiological outcomes and may alter experimental results making it difficult to replicate results between labs. Recent studies identified seven mutations in commonly used stock strains of E. coli K-12 MG1655 with implications for physiological experiments including loss of function of *glpR* and *crl* [5]. *C. difficile* is no exception to this phenomenon. Previous studies have demonstrated that accumulated mutations in stock strains can have physiological implications and even altered virulence in a hamster infection model [6]. Thus, with an explosion of research on *C. difficile* it is important to delineate mutations in stock strains and explain their physiological consequences.

To investigate the hypothesis that strains passaged in different laboratories would exhibit divergent phenotypes, we generated large scale metabolic profiles of carbon utilization for three isolates of a reference strain commonly used in *C. difficile* research: CD630 isolates from two different laboratories as well as one close relative sensitive to the antibiotic erythromycin

(CD630Δerm). Furthermore, whole genome sequencing of the strains allowed a comparison of both the genetic and phenotypic divergence amongst the three laboratory stock cultures. Genome-scale models (GEMs) of metabolism serve as a unifying platform to advance coordination of research and therapeutic advancements [7, 8]. To contextualize the divergence in phenotype and genotype between our stock strains we built and used a new genome-scale model of *C. difficile* 630.

GEMs offer a systems-level analysis of an organism's metabolic capabilities and establish a formal relationship between genotype and phenotype [9]. Two previous reconstructions iMLTC806cdf [10] and icdf834 [11] have been published for *C. difficile* strain 630. Here we present iCN900 that builds on iMLTC806cdf and icdf834, and reflects the most comprehensive knowledge base for *C. difficile* 630 to date. The model was used as a scaffold to interrogate the issue of stock culture evolution. We analyzed the core-genome of 415 strains to identify allelic sequence variants between genes determined to be present in each of the strains. Analyzing these genes within the metabolic network context provided by iCN900 illuminates which *C. difficile* metabolic pathways may be under evolutionary selective pressures. Additionally, these data emphasize how laboratory-specific microenvironmental pressures on stock cultures contribute to divergent interlaboratory results that may hinder translational science limiting the development of new treatment options.

4.3 Results

4.3.1 High-throughput screens highlight phenotypic differences between three CD630 lab strains

To evaluate the phenotypic divergence of closely related strains, we selected three different laboratory strains of *C. difficile* 630 including the close relative knockout 630Δerm strain [6]. We refer to the three strains as Savidge 630, Britton 630, and Britton 630Δerm coinciding with their laboratory of origin, noting that the Britton 630 strain is not parental to Britton 630Δerm (Methods). Phenotypic growth profiles of all three strains were generated in biological triplicate across 190 different carbon sources using Biolog Phenotype Microarrays [12]. Using the growth data generated from each *C. difficile* strain, we evaluated the phenotypic divergence of these closely related strains. Overall, each of the three laboratory *C. difficile* strains showed concordant phenotypes on 158 of the 190 compounds tested (Figure 4.1B). Thirty-two (16.8%) compounds displayed varied growth phenotypes across this set of 3 lab-adapted CD630 strains including several notable differences that are interrogated using the genome-scale model discussed below (Figure 4.1A).

To robustly evaluate the genetic content of each of our three investigated laboratory 630 strains, we completed whole genome sequencing and comparative genomics analyses to identify genetic differences relative to the reference 630 sequence (AM180355.1). We used breseq [13] to identify single-nucleotide variants (SNVs) and gene deletions with respect to the reference sequence (Figure 4.1C, 4.1D). Complete lists of predicted variants and deletions are available in Tables 4.1 and 4.2, respectively. Seven variants previously noted as likely mistakes in the original *C. difficile* 630 AM180355.1 reference assembly [6] were identified in all three strains. An addi-

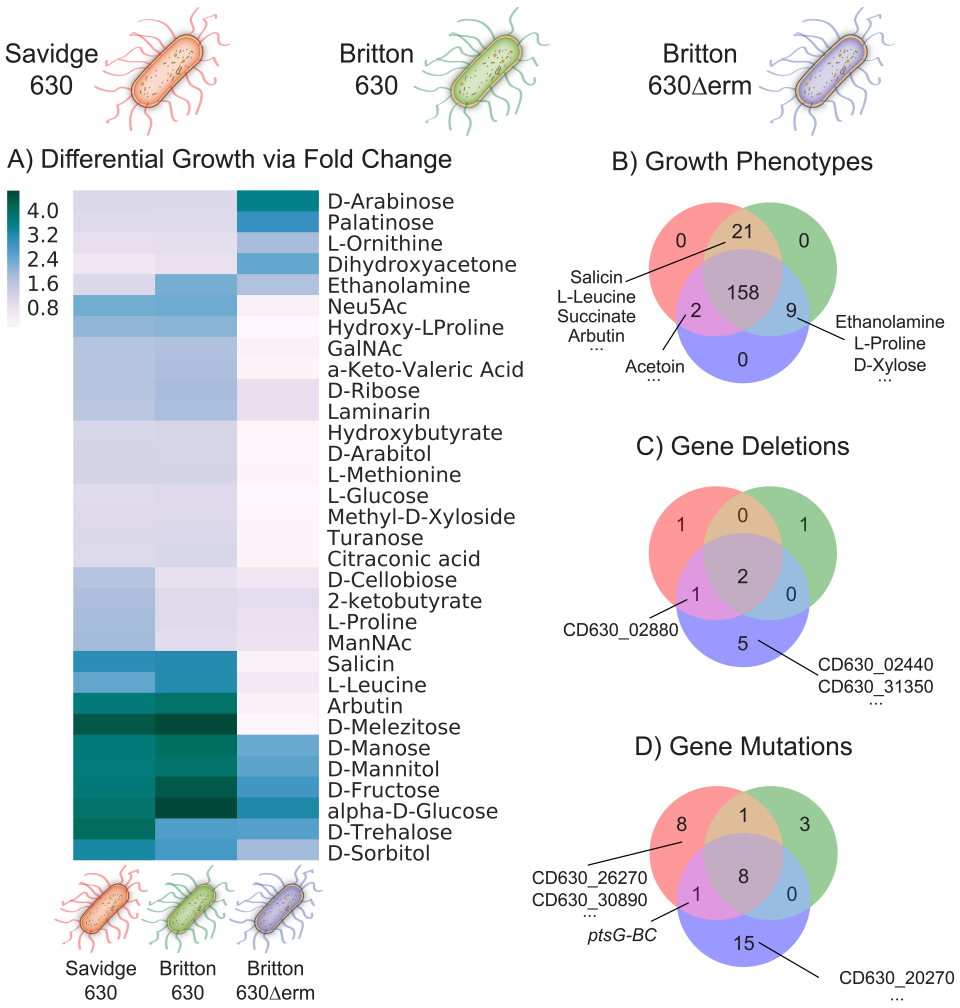

Figure 4.1: Experimental phenotyping of three different laboratory stock cultures of C. difficile 630. The Savidge 630, Britton 630, and Britton 630Δerm are represented by red, green, and blue respectively. A) Heat map of the maximal OD620 of C. difficile strains in Biolog phenotype microarray plates for which the fold change among the strains had the greatest standard deviation between the strains. Selected carbon substrates supporting differential fold change are shown. (n=3 biological replicates per strain). B) Venn diagram of 190 carbon substrates tested. All three strains shared 158 growth phenotypes, while 21 phenotypes were shared between Savidge 630 and Britton 630, 9 between Britton 630 and Britton 630Δerm, and 2 phenotypes between Britton 630Δerm and Savidge 630. C) Venn diagram detailing the identified gene deletions of each strain versus the reference sequence. D) Venn diagram detailing mutations of each strain versus the reference sequence.

tional synonymous SNV (E304E (GAG→GAA)) within the aminotransferase gene CD630_25320 was identified as common in all three strains. The Savidge 630, Britton 630, and Britton 630Δerm

each had 8, 3, and 15 unique SNVs relative to the reference (Table 4.1).

Table 4.1: Comparison of SNVs detected across three *C. difficile* 630 laboratory stock strains.

Gene	Mutation	Annotation	Savidge 630	Britton 630	630 Δerm
CD630_05730-thrS	C → T	intergenic (+173/-843)	Yes	Yes	Yes
CD630_05770-05780	A → T	intergenic (-126/+35)	Yes	Yes	Yes
CD630_24550-024560	G → T	intergenic(-543/-193)	Yes	Yes	Yes
rplC	G → T	G114G (GGG → GGT)	Yes	Yes	Yes
CD630_11900	T → C	F133L (TTT → CTT)	Yes	Yes	Yes
CD630_17670	C → G	P33A (CCC → GCC)	Yes	Yes	Yes
CD630_25320	C → T	E304E (GAG → GAA)	Yes	Yes	Yes
CD630_13880	(T) 6→7	coding (40/45 nt)	Yes	Yes	Yes
CD630_31561	+A	coding (309/339 nt)	Yes	—	Yes
CD630_34170-34180	A → G	intergenic (-3769/+1786)	Yes	Yes	—
CD630_34170-34180	+ C	intergenic(-3628/+1927)	Yes	—	—
CD630_19000-19010	A → T	intergenic (-160/-294)	Yes	—	—
CD630_02050	G → T	G165C (GGT → TGT)	Yes	—	—
CD630_26850	Δ21 bp	coding (339-359/1770 nt)	Yes	—	—
CD630_32450	C → T	E261K (GAA → AAA)	Yes	—	—
CD630_26270	C → A	G68C (GGT → TGT)	Yes	—	—
CD630_30890	T → G	E258D (GAA → GAC)	Yes	—	—
CD630_26670	C → T	V228I (GTT → ATT)	Yes	—	—
CD630_26670	A → C	*524E (TAA → GAA)	—	—	Yes
CD630_26670	(T) 8→7	coding (1558/1572 nt)	—	—	Yes
CD630_20270	G → A	G373E (GGG → GAG)	—	—	Yes
CD630_06430	T → C	I199I (ATT → ATC)	—	—	Yes
CD630_07610	G → T	D136Y (GAC → TAC)	—	—	Yes
CD630_12480	G → T	G59V (GGC → GTC)	—	—	Yes
CD630_14040	A → G	E536G (GAA → GGA)	—	—	Yes
CD630_12740	C → T	Q386* (CAA → TAA)	—	—	Yes
CD630_22630	G → T	S127* (TCA → TAA)	—	—	Yes
CD630_22670	(A) 5→6	coding (280/321 nt)	—	—	Yes
CD630_29430	T → C	N210D (AAT → GAT)	—	—	Yes
CD630_33790	C → A	E63D (GAG → GAT)	—	—	Yes
CD630_33980	C → A	G9C (GGT → TGT)	—	—	Yes
CD630_30360-30370	G → T	intergenic (-1521/+386)	—	—	Yes
treR	Δ6 bp	coding (192-197/723 nt)	—	—	Yes
CD630_12060	A → T	K120N (AAA → AAT)	—	Yes	—
CD630_27920	T → A	P669P (CCA → CCT)	—	Yes	—
CD630_31840-31850	Δ9 bp	intergenic (-393/+222)	—	Yes	—

Two gene deletions were identified common to all three genomes and a third deletion was present only in 630 Savidge and Britton 630Δerm. Savidge 630 and Britton 630 each contained a single independent deletion relative to the reference in conserved hypothetical proteins CD630_01960 and CD630_12100 respectively. In contrast, Britton 630Δerm contained five unique deletions not present in Savidge or Britton 630 (See Table 4.2). Three of the deletions unique to Britton 630Δerm are two groups of 44 and 46 genes annotated as putative phage genes and the expected 8 gene loss for the erythromycin-sensitive derivative. In order to contextualize the remaining genetic differences distinguishing these three strains from each other and their impact on the observed phenotypic divergence we updated and deployed a genome-scale model of *C. difficile* 630 metabolism.

Table 4.2: Comparison of deletions detected across three *C. difficile* 630 laboratory stock strains.

Gene	Description	Savidge 630	Britton 630	630 Δerm
CD630_10250	ABC-type transport, spermidine	Yes	Yes	Yes
CD630_34170-34180	ABC-type transport, sugar-family	Yes	Yes	Yes
CD630_02880	PTS system, mannose/fructose	Yes	—	Yes
CD630_01960	conserved hypothetical protein	Yes	—	—
CD630_12100	conserved hypothetical protein	—	Yes	—
CD630_02440-02450	CDP-glycerophosphotransferase	—	—	Yes
CD630_31350	Fructose-1-6-biphosphate adolase	—	—	Yes
CD630_09390-09770	46 genes: putative phage protein	—	—	Yes
CD630_28900-29521	44 genes: putative phage protein	—	—	Yes
CD630_20060-ermB1	8 genes	—	—	Yes

4.3.2 Genome-scale network reconstructions contextualize genetic divergence by serving as a scaffold for structural systems biology analysis

Genome-scale models offer a powerful tool to contextualize and explain the effect of genetic changes in pathogenic organisms that impact human health. Therefore, we evaluated

and updated a genome-scale network reconstruction of *C. difficile* 630. The new *C. difficile* GEM, iCN900, contains an additional 66 genes, 46 reactions, and 70 metabolites compared to previous models of this strain. New content was incorporated into the reconstruction using both bioinformatic tools and manual curation. We implemented several tools to add new content to the reconstructions including the enzyme detection tool DETECT v2 [14], searching for homologs in closely related reconstructions, [15], and manual curation of pathways based on false negative model predictions against experimental data (Methods). This allowed for the inclusion of new transport reactions as well as significant refinement of the accuracy of the gene product rules for existing transporters. GEM additions included reactions for tRNA synthetase, carbon and sulfur metabolism, and cell envelope biosynthesis.

In addition to adding new content to the genome-scale network reconstruction for *C. difficile* 630, another major area of improvement was the removal of erroneous energy generating cycles (EGCs). EGCs allow for free energy generation during flux balance analysis (FBA) simulations and have been shown to be a prevalent problem in many non-curated GEM predictions. We implemented an existing algorithm [16, 17] to identify and confirm the existence of EGCs in the previous model (icdf834). We manually investigated icdf834 and found erroneous EGCs for ten energy carrying metabolites. We edited the reversibility of 29 reactions to remedy these cycles making the network completely devoid of EGCs and therefore better suited to make accurate flux predictions utilizing FBA [18]. Finally, we updated the model nomenclature to align it with the BiGG standard, making its contents directly comparable with over 100 reconstructions of diverse organisms present in the BiGG database [19, 20]. This improved model, iCN900, is available in the BiGG database.

Recent studies have supplemented GEMs with protein structures to form GEM-PROs

resulting in expanded applications for both genome- and protein-scale models [21, 22]. This approach has enabled further contextualization of SNVs within the metabolic network. Protein structures have never been incorporated with a GEM of *C. difficile*, therefore we evaluated the current state of structural data available for *C. difficile* by mapping protein structures to the Protein Data Bank (PDB). Overall 1,221 genes within the 630 reference genome map to a structure within the PDB. A subset of 524 of these genes are contained within iCN900. However, only 2.5% (29/1,145) of mapped structures with less than 75 percent identity (PID) are sourced from *C. difficile*. Conversely, 85.5% (65/76) of mapped structures with greater than 75 PID are *C. difficile* specific (Supplementary Figure C.1). This steep drop off in the number of *C. difficile* mapped structures demonstrates the overall structural knowledge gap for *C. difficile*. Only 20 of the genes within iCN900 map to a structure that is greater than 75 PID and sourced from *C. difficile*. These represent the best characterized, metabolically related *C. difficile* specific structures [23–26].

4.3.3 Experimental validation of iCN900 demonstrates high model accuracy

We evaluated iCN900 by performing simulations on four in silico media types as delineated by Larocque et al [10, 11]: 1) Minimal, 2) Basal Defined Medium (BDM), 3) Complete Amino Acid-Defined medium (CADM) and 4) Complex media. We confirmed biomass production by iCN900 under each in silico media type and further showed that flux through the objective function increased commensurate with the complexity of media type (Figure 4.2A). We also confirmed that known essential amino acids required by *C. difficile* growth (cysteine, leucine, isoleucine, proline, tryptophan, and valine) [27, 28] are also required for biomass production.

To assess gene-essentiality prediction by iCN900, we performed in silico single gene dele-

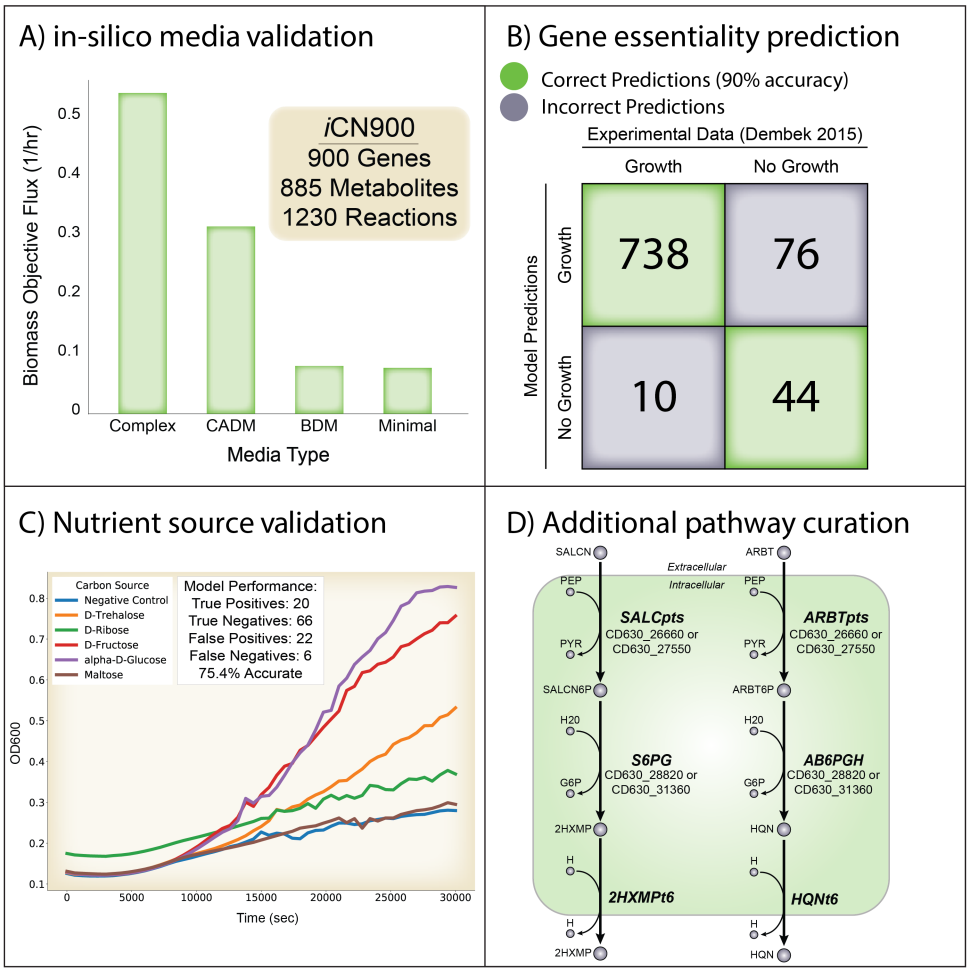

Figure 4.2: Properties and validation metrics of iCN900 A) Model predictions for biomass flux on four different in silico media types: Complex media, CADM, BDM and minimal media. Importantly, the biomass objective flux reflects the increasing amount of nutrients in each media condition. The overall gene, reaction, and metabolite content of iCN900 is summarized within the inset box. B) Comparison of model predictions of essential genes on complex media compared to experimental gene-knockout results from Dembek et al. C) *C. difficile* optical density at 620 nm was measured over time in Biolog Phenotype Microarray plates. Representative growth curves for the Savidge 630 strain on 5 indicated carbon sources (of the 190 tested) and the negative control are shown. Experimental growth of *C. difficile* was compared to iCN900 metabolic flux predictions, to determine the accuracy of predictions as summarized in the inset box. D) Putative metabolic pathways for *C. difficile* utilization of salicin and arbutin were incorporated into iCN900 through targeted gap-filling enabled by comparison to experimental growth data.

tions and compared these predictions to an available experimental dataset of essential genes for *C. difficile* in strain R20291 [29]. *C. difficile* R20291 is evolutionarily distinct from strain 630

and the iCN900 model achieved an overall accuracy of 90% for these gene-essentiality predictions (Figure 4.2B). The true negative gene predictions are predominantly associated with reactions encoding lipid metabolism indicating that in complex media this portion of the metabolic network is particularly sensitive to single gene knockouts both in silico and in vitro. Examining the 10 false negative predictions revealed genes involved in pyrimidine metabolism indicating that perhaps R20291 has alternative encoding mechanisms for reactions in this pathway. Future improvements to CD630 reconstructions would benefit from experimentally validated gene essentiality datasets specific to this strain.

Finally, we validated the ability of the iCN900 model to predict growth capabilities on 190 diverse carbon sources by comparing model predictions to the phenotypic microarray growth data generated for the three independent laboratory strains of *C. difficile* 630 (Figure 4.2C). In silico growth predictions of the iCN900 model were generated using previously defined minimal media conditions (Methods) and alternating the carbon source to coordinate with that being tested in the experimental microarray [10, 27]. Of the 190 carbon sources tested, 114 were represented in the model and overall model predictions agreed with experimental growth capabilities for 75.4% of cases.

4.3.4 Targeted gap-filling of incorrect model predictions uncovers new catabolic pathways in *C. difficile* metabolism

Comparison of phenotypic screens to model predictions can be used to iteratively improve genome-scale model reconstructions by informing the inclusion of metabolic pathways missing in the network content. Using the phenotypic microarray growth data generated from each *C. difficile* strain, we evaluated the phenotypic divergence of closely related strains against our

curated iCN900 genome-scale model. Our data confirmed previously published studies [30–32] and verified growth of two of the three *C. difficile* 630 strains on salicin, arbutin, and N-acetyl-galactosamine (GalNAc) (Figure 4.1A). However, initially the iCN900 model predicted the inability of CD630 to grow on these compounds. Both salicin and arbutin are β-glucosides and are produced in various plant species thus it is plausible that these compounds could be available within the human gut dependent on diet [33, 34]. We identified homologous genes in the pathways for catabolism of these two compounds in *Bacillus subtilis*, a close relative of *C. difficile*. Our identified candidate pathways have a similar pathway architecture: a transporter (encoded for by *ptsG-A* and *ptsI*), a glucohydrolase (encoded for by *celF* and *bglA7*), and efflux of 2-hydroxymethyl-phenol or hydroquinone respectively, both products of the respective glucohydrolase. Homologs in the *C. difficile* 630 genome were identified and incorporated into iCN900 using gene product rules based on homology with B. subtilis (Figure 4.2D) and the experimental evidence that these compounds support growth.

Like salicin and arbutin, our experimental growth assays verified N-acetyl-galactosamine was sufficient to support growth of two of the three *C. difficile* 630 strains tested, but this phenotype was absent from our initial rendition of the iCN900 model. N-acetyl-galactosamine is of particular interest because as a host-derived glycan it is proposed to be an important carbon and nitrogen source for *C. difficile* in the gastrointestinal tract [30]. We hypothesized that N-acetyl-galactosamine utilization would be facilitated by a phosphotransferase system (PTS) similar to those seen in other enteric bacteria and investigated other GEMs for N-acetyl-galactosamine catabolic pathways. We identified an isomerase encoded by agaI in E. coli [35] that converts N-acetyl-galactosamine-6-phosphate to tagatose-6-phosphate. Our experimental dataset indicates all three *C. difficile* 630 strains grew significantly (P=.006, Paired T-Test) in the presence

of tagatose (6.59 fold-change relative to negative control) but did not grow on galactose (0.89 fold-change), thus supporting the possibility of this interconversion. In agreement with the experimental results, iCN900 predicts *C. difficile* growth on tagatose (true positive), and no growth on galactose (true negative). This inference along with the strength of the experimental evidence led to the inclusion of the PTS and isomerase within iCN900. Further experimental work to identify any additional genes that encode this machinery would increase understanding of N-acetyl-galactosamine utilization by *C. difficile*, which may have important implications in the context of infection.

Surprisingly, iCN900 predicted an inability to be grown in ethanolamine, which is in contrast to our experimental evidence and the literature that many gut bacteria, including Clostridia, are capable of ethanolamine catabolism as a sole carbon or nitrogen source [36]. Furthermore, phosphatidylethanolamine is a prevalent membrane phospholipid, which is catabolized into glycerol and ethanolamine, suggesting that ethanolamine is an abundant nutrient in the gastrointestinal tract. iCN900 contains the genes of the eutG operon and the corresponding enzymes for usage of ethanolamine [37]. Previous studies have shown that *C. difficile* 630 strains can utilize ethanolamine in vitro, however the media conditions in these studies included glucose along with ethanolamine [37]. We postulated that if phosphatidylethanolamine is a primary source of ethanolamine within the gut, then glycerol would be concurrently available to *C. difficile*. Interestingly, glycerol scored as a false positive in our initial prediction. Further analysis revealed that when both glycerol and ethanolamine were components of the in silico minimal media the biomass objective flux increased to .034 from .014 on glycerol alone or 0 on ethanolamine alone. This apparent synergistic usage predicted by iCN900 of these two metabolites is interesting given their likely co-availability in the host. Glycerol as a sole carbon source has a limited

uptake flux value of 4.56 and valine and leucine were identified as non-carbon limiting nutrients. When both ethanolamine and glycerol are available both have an uptake flux of 10, indicating an energetically favorable complement of catabolic pathways. Ethanolamine utilization produces acetyl-CoA which is a key metabolite in many downstream metabolic pathways. We hypothesize that the ability to use ethanolamine as a source to produce the necessary acyl-carrier proteins frees glycerol to be used for other growth requirements. While no modifications were made to the network to change the determination of glycerol as a false positive prediction and ethanolamine as a false negative prediction, it is worth noting this potential feature of *C. difficile* physiology and a future validation of this prediction would be valuable.

4.3.5 iCN900 links observed mutations to unique phenotypes

With an updated reconstruction completed, we used this resource to evaluate the mutations and deletions observed between the three reference strains (Table 4.1 and 4.2). Of the deleted genes, two are implicated in the metabolism of fructose and mannose that are of particular interest. First, Savidge 630 and Britton 630Δerm each contain a deletion of CD630_02880, which is part of the GPRs for both fructose and mannose PTS reactions. Secondly, a unique deletion in Britton 630Δerm of CD630_31350, a gene involved in the fructose bisphosphate aldolase reactions. Growth results reveal the maximum fold change in optical density during fructose utilization is 24.3% lower (P=0.24, Paired T-Test) in Britton 630Δerm versus Savidge 630, and 34.8% lower (P=0.008) versus Britton 630. During mannose utilization, growth reduction is 35.1% (P=0.1) and 40% (P=0.03) respectively. While there is no significant decrease in growth on both sugars between Britton 630Δerm and the Savidge strain, the decreases between Britton 630Δerm and Britton 630 are both statistically significant. Given the co-occurrence of deletions

in the transport systems for these sugars and fructose bisphosphate aldolase reactions, we hypothesize that the deletions together result in the observed growth reduction for the Britton 630Δerm strain with perhaps the more consequential deletion being CD630_31350.

Mutations within coding sequences and particularly those in genes annotated with metabolic functions were prioritized. The Savidge 630 strain possesses a substitution in the aspartate kinase gene (G68C (GGT→TGT)). However, there were no physiological changes in the growth experiments on aspartic acid, which is likely explained by the presence of aspartic acid in the basal medium. Savidge 630 also contained a unique nonsynonymous substitution (E258D (GAA→GAC)) in CD630_30890, which is part of the gene product rule for the trehalose phosphotransferase reaction. Analysis of the growth screen data indicated that the maximal optical density of the Savidge 630 strain during trehalose utilization was over 30% greater (P=.04) than either the Britton 630 or the Britton 630Δerm strains. Mapping of this substitution to the predicted protein structure reveals that it occurs within a hydrophilic region of the protein (Figure 4.3A), suggesting that the substitution may confer an advantage to the import or phosphorylation of trehalose entering the cell. To test this hypothesis, growth curves in minimal medium supplemented with 10 mM, 25 mM, and 100 mM trehalose were compared (Figure 4.3B), revealing that at the lower concentration of trehalose the Savidge 630 strain grew significantly better than the other two strains ($P \leq 0.0001$). However, in higher concentrations of trehalose, the growth of the Britton CD630Δerm isolate (25 mM) and the Britton 630 isolate (100 mM) matched that of Savidge 630. The significant increase in growth at 10mM supplementation of trehalose indicates that the substitution may increase affinity of the PTS for trehalose transport, improve efficiency of transport, or increase expression and/or stability of the transporter however the role of the this E258D substitution in trehalose uptake still needs to be explored. Analysis

of the growth curves by Gaussian Process modeling [38] allowed us to quantify growth rate, area under the curve, and carrying capacity of the isolates in each condition (Figure 4.3C).

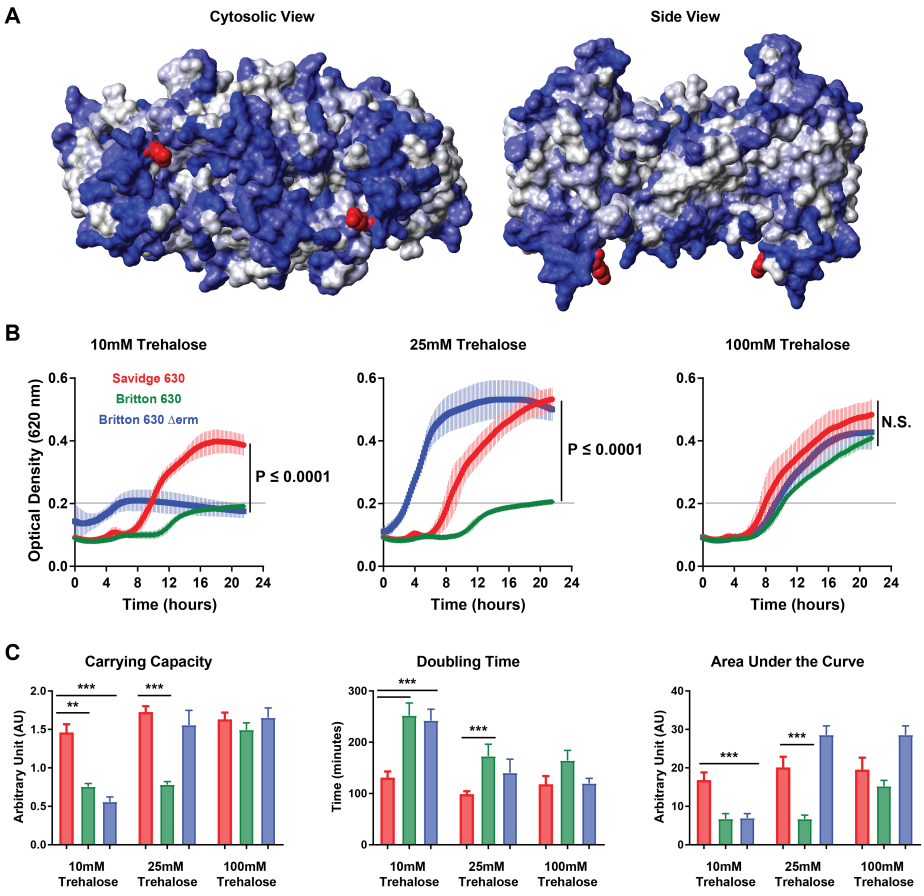

Figure 4.3: Characterization of phenotypic growth differences of lab adapted isolates on trehalose A) Predicted protein structure of *C. difficile* PTS (CD630_30890) based upon the crystal structure of the MalT transporter. The EIIC domain is shown as a dimer, with the E285D substitution of the 630-Savage isolate highlighted in red on the cytoplasmic interface. The model shading indicates amino acid hydrophobicity (gray residues are hydrophobic and blue residues are hydrophilic according to the Kyte-Dolittle scale). B) Growth curves of *C. difficile* isolates, Savidge 630 (red), Britton 630 (green) and Britton 630Δerm (blue) in defined minimal medium supplemented with trehalose. The gray line indicates the maximal optical density of the negative control wells. Optical density at 620 nm measured at 10 minute intervals, The plotted bar is the mean of 3 biological replicates assayed in duplicate wells and the error bars represent the standard deviation of the mean. C) Growth curves from the conditions in (B) were analyzed by Gaussian process curve fitting to calculate the total carrying capacity, doubling time, and total area under the curve (error bars represent the standard deviation of the mean ** = P≤0.001, *** = P≤0.0001).

Both the Savidge 630 strain and the Britton 630Δerm strain had unique mutations within the CD630_26670 gene, which codes for part of the PTS reaction for α-glucose. In the Savidge 630 strain the single nucleotide polymorphism results in a substitution of isoleucine for valine (V228I (GTT→ATT)), however the mutation in the Britton 630Δerm strain switches the stop codon to glutamic acid (*524E (TAA→GAA)). The loss of this stop codon in the Britton 630Δerm strain results in extension of the CD630_26670 coding region directly into the downstream gene with the next stop codon at position 691. The lack of a stop codon would likely produce an aberrant transcript subject to degradation by cellular regulatory mechanisms [39]. Analysis of the growth data supports the hypothesis these substitutions impair the import of α-glucose as growth via maximum optical density of the Savidge 630 strain is reduced by 21.7% and the Britton 630Δerm strain is reduced by 30.4% compared to Britton 630 strain (P=.05), which is devoid of any mutations in these genes. Overall, the mutational analysis provides insight into unintentional evolution occurring in laboratory strains and highlights the need for resequencing strains used commonly across many labs to more accurately reflect the heterogeneity among reference sequences. This is particularly important for the accuracy of corresponding genome scale models and downstream constraints-based analyses.

4.3.6 iCN900 applied to analyze sequence variation within the *C. difficile* core-genome

We used the iCN900 model to link mutations amongst the three strains to the differences observed within the phenotypic growth profiles. iCN900 is specific to *C. difficile* 630, one of the most well characterized strains and often used as a reference strain in studies. However, we have shown that there is genetic divergence within even 630 stock cultures from different laborato-

ries. As demonstrated above, single nucleotide variations can manifest themselves as deviations in metabolic profiles pointing to the importance of even small amounts of genetic divergence between *C. difficile* isolates. Therefore it is worth considering the sequence variation amongst shared genes within several strains of the species. To this end we used bi-directional BLAST to identify the genes within *C. difficile* 630 present at greater than 80 PID in 415 high-quality, publicly available genomes (Figure 4.4A). From these genes, those that were present in more than 99% (411/415) of the strains were determined to comprise the core-genome of *C. difficile*. A total of 2,756 of 3,828 *C. difficile* 630 genes comprise the core-genome. iCN900 was then utilized to investigate the metabolic core-genome which consisted of 765 core metabolic genes. A genome-scale model based on the function of these 765 genes was created to investigate core *C. difficile* metabolic capabilities; iCN765. This representation of the core metabolic functions of the *C. difficile* species represents a potentially valuable starting point for reconstruction of other strains. The core model was used to investigate metabolic phenotypes common to all strains of *C. difficile*. Simulations with in-silico minimal media predict that the core metabolic network cannot produce biomass. However, media supplementations were identified that enable synthesis of certain biomass constituents. Protein synthesis required supplementation with histidine, lysine, arginine, and threonine. Supplementation with uridine or uracil enabled DNA and RNA synthesis and nicotinate supplementation enabled associated cofactor production. Following these media supplementations the core network still lacks the ability to produce the lipid and peptidoglycan biomass components. Performing gene essentiality analysis on the full *C. difficile* 630 model using this supplemented in-silico media condition predicts that there are 4 non-core genes which are essential for the production of lipids and peptidoglycan. Upon further examination, 3 of these genes are present within 96% (398/415) of the strains, thereby designated as non-core and

perhaps the strains without these genes have either acquired alternative encoding mechanisms or vary in lipid/peptidoglycan composition. It is worth noting that the strains without these 3 genes represent the strains of type MLST11 and MLST254 within the group of 415. The final non-core gene essential to production of peptidoglycan is only present within 13% (54/415) of strains and is involved in the production of teichoic acid for cell wall synthesis.

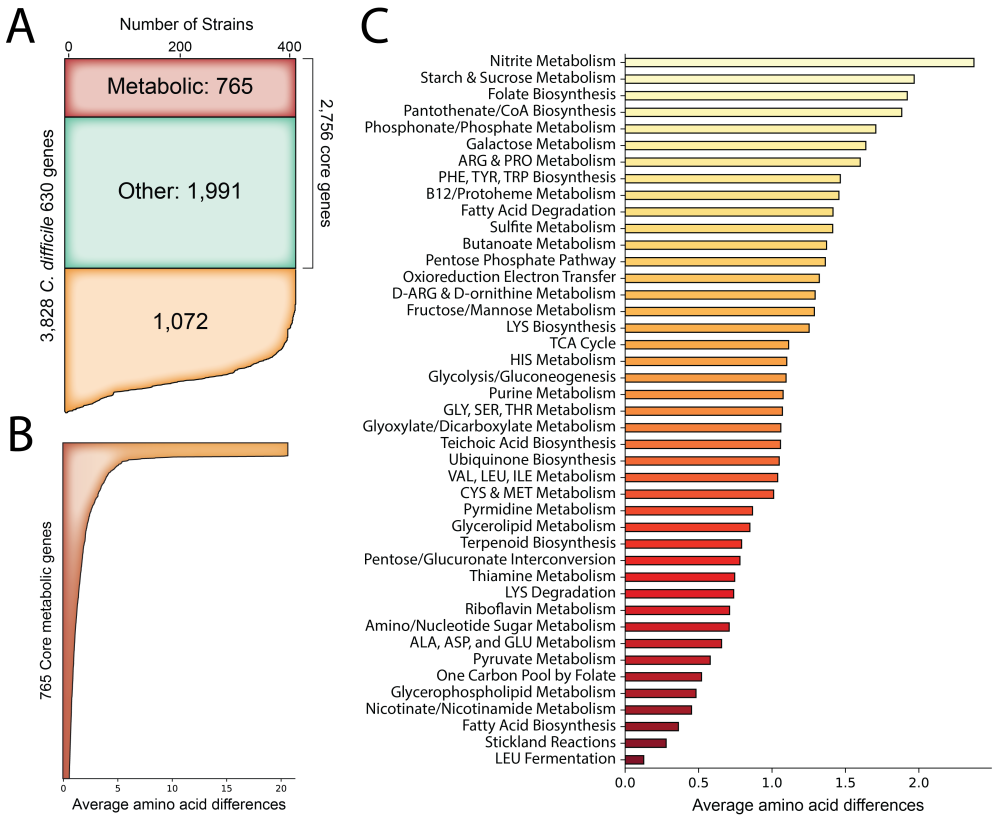

Figure 4.4: Core-genome of *C. difficile* reveals metabolic subsystems with greater sequence variation A) By comparing the genomes of 415 publicly available *C. difficile* genomes the core-genome was calculated and includes 765 metabolic genes. B) Analyzing the sequence variation among the 765 core metabolic genes demonstrates that the average difference in amino acid sequence range from 0 to just over 20 for these shared genes. C) The genome scale reconstruction enables stratification of the genes by metabolic subsystem and comparison of average amino acid differences of each gene within a subsystem. This reveals that nitrite and starch/sucrose metabolism have the highest degree of sequence variation whereas Stickland reactions and leucine fermentation are the most conserved.

Beyond investigating conserved metabolic functions, examining the conserved sequences amongst the 415 strains provides other novel insights (Figure 4.4B). In the core metabolic gene products, we evaluated the average amino acid difference and found them to range from zero (completely conserved amino acid sequence across all strains) to just over 20 average amino acid differences between strains. For example, we identified strain FDAARGOS_268 (PATRIC ID:1496.2022) with the same trehalose phosphotransferase (CD630_30890) E258D mutation described in Savidge 630 above as well as strain QCD-32g58 (PATRIC ID: 367459.5) with an E258K substitution in the same protein. Strain QCD-32g58 was isolated in 2017 from a patient in Quebec, Canada with severe *C. difficile* infection and is noted to be a representative of a predominant Quebec strain. Furthermore, the greatest average amino acid differences (>20 average amino acid differences) occurred in two gene products, CD630_01370 and CD630_35270, that are implicated in transport reactions for cellobiose and iron, respectively. Each individual gene can also be interrogated for the frequency of each allele sequence within the group of 415 strains (Figure 4.5A) and these sequences can be compared for their similarity to one another (Figure 4.5B). For the genes that are part of the GEM-PRO the mutations per allele can be mapped to the representative structure providing a three dimensional view of the effect of the change (Figure 4.5C). We performed this analysis for the *thiD* gene encoding phosphomethylpyrimidine kinase and gained insight into the areas of the protein structure where the sequence variants manifested.

The GEM-PRO also allows for a systems level analysis of the variation within these core-gene products by stratifying the average amino acid differences per reaction to metabolic subsystems (Figure 4.4C). This network context illuminates the metabolic subsystems that may be under evolutionary selective pressures due to higher degrees of sequence variation. The reactions for nitrite metabolism, starch and sucrose metabolism, and folate biosynthesis have the

greatest variation indicating these are potential evolutionary hot spots. Conversely, leucine fermentation and Stickland reactions are the most conserved in terms of sequence suggesting that these enzymes and related functions are defining traits within the species.

To increase the analysis of metabolic network areas that may be under selective pressure within *C. difficile*, we considered the classification of enzyme specificity. Generally, it is understood that specificity is an evolutionarily beneficial trait towards increased catalytic efficiency. We used iCN900 to classify the genes and reactions within as either generalist or specialist. As previously reported [40] we define a specialist gene as one that participates in only one reaction and generalists as those involved in multiple reactions. We applied this criteria to all metabolic enzymes within and showed that there are 410 specialist genes encoding proteins catalyzing 287 specialist reactions and 231 generalist genes encoding proteins catalyzing 484 reactions. This distribution is similar to that previously found for *E. coli* [40]. Of the specialist genes, 76 encode subunits of a complex and 148 are isozymes. Similar to our analysis of the sequence variation of the core-genome, we used the reconstruction to evaluate the distribution of specialist and generalist reactions per metabolic subsystem. Analyzing each subsystem we found that certain subsystems were enriched in specialist enzymes and others in generalist enzymes. Starch and sucrose metabolism, folate biosynthesis, vitamin B12 and protoheme metabolism, and histidine metabolism are all enriched in specialist reactions (hypergeometric $P<0.05$). Valine, leucine and isoleucine metabolism, glycerolipid metabolism, one carbon pool by folate, and fatty acid biosynthesis are all enriched in generalist reactions (hypergeometric $P<0.05$). Consistent with the calculation of sequence variation amongst subsystems the specialist enriched subsystems had an average of 1.61 amino acid differences and the generalist enriched subsystems had an average of .69 average amino acid differences (Supplementary Figure C.2). These network based analyses

enabled by the reconstruction provide insights into the pressure surrounding the core metabolism of *C. difficile* as a species and point to vulnerable processes worth investigating as potential drug targets.

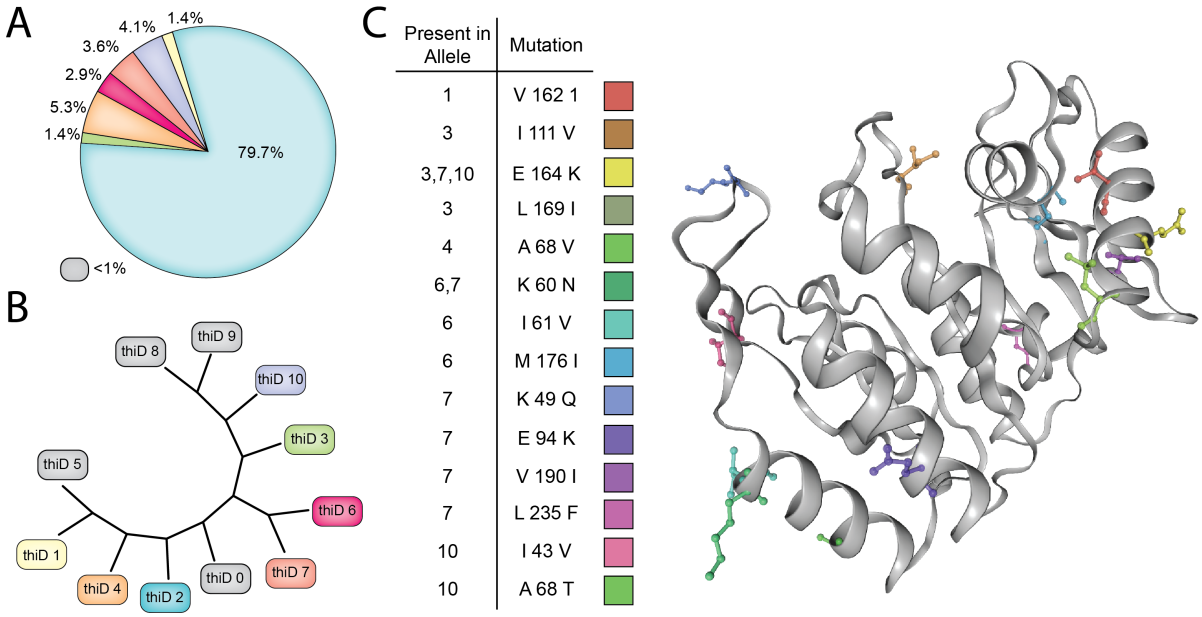

Figure 4.5: Allele diversity for *thiD* as an example of sequence diversity. A) The 415 sequences for the thiD gene have 11 variant sequences (alleles) variably present within the population. Notably the reference sequence allele is present within 79.7% of the population whereas the next most frequent allele is present in 5.3% of the population. B) The degree of similarity between each sequence is readily accessible. For example the *thiD* 6 and *thiD* 7 sequences are similar to one another sharing a K60N mutation. C) Through the use of the GEM-PRO each mutation by variant can be visualized within the 3D space of crystal structures where applicable.

4.4 Discussion

Genome scale metabolic network reconstructions provide a valuable format to unify disparate knowledge about an organism, and contribute a tool that may be used to investigate an organism's properties. We developed the most comprehensive knowledge base for *C. difficile*

strain 630 to date and utilized the model to (i) investigate catabolic capabilities in conjunction with experimental data; (ii) serve as a framework for investigation into genetic drift amongst different laboratory *C. difficile* 630 strains and a derivative strain; (iii) analyze the sequence variation amongst the genes within the core-genome of *C. difficile*. The GEM performs with as much as 90% accuracy in predicting gene essentiality and 75% accuracy in predicting catabolic capabilities. The metabolic network represented within iCN900 is devoid of EGCs and the standardization of reaction and metabolite identifiers opens up the possibility of inclusion in studies of multiple organisms that share this namespace. Phenotypic profiling and model driven discovery identified new pathways potentially relevant to *C. difficile* survival due to their presence in the diet (arbutin and salicin) or as components of the human gut (N-acetyl-galactosamine). By coupling the generation of the new reconstruction, iCN900, with extensive phenotypic profiling and further genome analytics we have increased the body of knowledge about this pathogen.

The process of crafting iCN900 evoked questions of genetic drift amongst isolates of the same strain of bacteria. The variability in both genotype and phenotype of isolates that are either deemed strain 630 or are closely related points to the need to resequence strains used in experiments and to recognize that reference sequences represent only a single time-point in the lifetime of a strain. This point was borne out in our comparison of the trehalose transporter between laboratory strains. Hypervirulent strains of *C. difficile* are known to metabolize trehalose, a process recently attributed to hypervirulent strain evolution coinciding with the widespread adoption of trehalose in our diet [41] Microevolution of strains within laboratories could impart divergent conclusions between laboratories undergoing similar experimental processes to evaluate pathogen evolution and virulence, which may serve to hinder translational science and limit new treatment options. This phenomena has been observed in other model organisms including

E. coli [5] and yeast [42]. In *E. coli*, *glpR* mutations have been observed leading to constitutive expression of genes involved in glycerol catabolism likely due to repeated passage on glycerol containing media. Similar unexpected *glpR* alleles have been found in several other *E. coli* strains [43]. Thus a similar process of unintentional domestication of laboratory *C. difficile* strains based on adaptation to laboratory media may be underway. Given the importance of metabolism in infection kinetics and virulence, diligence in tracking genetic drift within strains will collectively improve scientific rigor and reproducibility with the potential to strengthen bodies of scientific evidence between laboratories.

Motivated by the demonstrated divergence in metabolic profile from small amounts of genetic diversity, the core-genome of *C. difficile* was constructed based on 415 publicly available genome sequences and sequence variation was analyzed. he reconstruction was used to identify metabolic traits common to the species and amino acid differences and enzyme specificity were used to evaluate which pieces of the metabolic network may be under selection pressures and those that are more conserved. Interestingly, and in agreement with the growing literature [41, 44, 45] concerning sugar metabolism of pathogenic *C. difficile* strains this analysis revealed that even conserved starch and sucrose metabolism genes are some of the most varied in terms of sequence. This demonstrates that *C. difficile* strains are actively evolving more efficient machinery to best adapt to their nutrient niche (be it in a lab or in the colon) and that unique catabolic capabilities could arise in response to availability of certain nutrients.

The generation of this high quality reconstruction enables future studies extrapolating this model across multiple strains to investigate species diversity. While we focused on core metabolic capabilities in this study, the exploration of accessory metabolic gene sets are underway and could give insight into the metabolic capacity specific to hypervirulent strain families of *C. difficile*. The

ability to identify evolutionary hotspots and specialized enzymatic reactions within hypervirulent strains may help direct drug development targeting previously unappreciated metabolic processes critical to pathogen survival. Furthermore, the ability to simulate coordinated changes in dietary supplements and predicted evolutionary hotspots could give insight into pathogen emergence.

4.5 Materials and Methods

4.5.1 Reconstruction

We began the reconstruction of iCN900 by using previous efforts iMLTC806cdf [10] and icdf834 [11] for *C. difficile* strain 630. This starting point was refined and translated to a reconstruction within the standardized BiGG namespace. This reconstruction was then extensively manually curated. Additionally, evaluation metrics as delineated in a protocol for generating reconstructions were executed [9]. Model content was iteratively improved by comparison to existing and generated experimental data. iCN900 reflects the final version of this iterative workflow.

4.5.2 Constraint-based Modeling

Constraints-based analyses were conducted using the COBRApy toolbox. For the in silico growth simulation of sole carbon source utilization the minimal media [27] was used and glucose was removed in an iterative fashion and other carbon source exchange reactions were opened to evaluate if growth was possible. Growth versus no growth was determined through flux balance analysis in each condition, optimizing for the biomass function. Within these simulations we consider biomass objective flux of greater than zero designated carbon sources that supported growth.

4.5.3 Protein Structure Integration

The GEM-PRO [21, 22] pipeline was used to annotate iCN900 with available protein structure information. The list of genes within iCN900 was mapped to sequences within Uniprot and consequently the Uniprot ID enables automatic mapping to the Protein Data Bank (PDB). The representative sequences are then BLASTed to the PDB and the best ranking structure available was identified for each model gene was identified and the quality of those rankings are presented.

4.5.4 Core-genome

A total of 1,246 whole-genome sequences of *C. difficile* were downloaded from the PATRIC database [46] on August 25, 2019. To filter for high-quality genomes a cutoff of assemblies composed of 100 or fewer contigs was applied. Furthermore, an MLST analysis of the genomes was performed using MLST [47, 48]. All genomes that could not be assigned to an MLST type or species were also filtered out. This led to a final set of 415 genome sequences for downstream analysis.

4.5.5 Designation of specialist and generalist enzymes

We classified 697 metabolic enzymes within iCN900 as either specialists or generalists. The selection criteria was a simplified approach as presented within [40] as the supplementary information to refine the approach is not as well defined for *C. difficile* as for E. coli. The 697 genes to be classified were selected from the reconstruction on the basis that they are not involved in any transport reactions. Following the definition of the group each was classified according to the following rule; specialist if the gene is present within the GPR of only one reaction and

generalist for those involved in more than one reaction. In turn it was possible to classify the encoded reactions in a corresponding manner as either specialist or generalist. The reaction classifications were then analyzed according to their metabolic subsystems and each subsystem was tested for enrichment of either class through the hypergeometric test.

4.5.6 Whole Genome Sequencing

Cryofrozen isolates of each *C. difficile* strain were incubated on Brain Heart Infusion (BHI) agar under anaerobic conditions for 24-48 h. Genomic DNA was extracted using the MasterPure Complete DNA RNA Purification kit (Lucigen, MC85200) and libraries of fragmented genomic DNA were prepared using NEXTflex Rapid DNA-Seq Kit (Bioo Scientific, NOVA-5149-02). Paired-end reads (2 x 150 bp reads) were generated on the MiSeq platform (Illumina, San Diego, CA, USA) using the Illumina MiSeq Reagent Kit v2 (MS-102-2002) and PhiX Control Kit v3 (FC-110-3001). Breseq v0.31 [13] was run with default parameters on each set of paired-end reads with the C. difficle 630 genome (AM180355.1) as a reference. We note that the individual CD630 strains utilized within this study have each been subcultured within their respective labs over time. The Britton 630 strain was received from a colleague at Tufts University on July 23, 2008 and the Savidge 630 strain was received from a colleague at the University of Houston in August 2014. Further we note that the Britton 630 strain is not the parent strain to Britton 630Δerm.

4.5.7 Phenotypic Profiling by Biolog

Strains were cultured in BHI medium (Difco) supplemented with 0.5% (w/v) yeast extract (Fischer Scientific) overnight (16 hours) in an anaerobic chamber (5% hydrogen, 90% nitrogen,

5% carbon dioxide). 1 ml of overnight culture was diluted into 10 ml of defined minimal media with previously described composition (Theriot et al, 2017) and 100 µl was added to each well of Biolog Phenotypic Microarray plates (PM1 and PM2). Growth assays were performed under anaerobic conditions with optical density at 620 nm read every 10 minutes over a period of 16 hours, in triplicate for each *C. difficile* 630 strain. Statistical analysis was performed by Two-way ANOVA, (with Tukey's correction for multiple comparisons where appropriate) in GraphPad Prism Software (v. 7.04).

Acknowledgements

Chapter 4, in part, is a reprint of material published in: **Norsigian CJ**, Danhof HA, Brand CK, Oezguen N, Midani FS, Palsson BO, Savidge TC, Britton RA, Spinler JK, Monk JM. "Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence." *NPJ systems biology and applications.* (2020): 6. The dissertation author was the primary author.

4.6 References

1. Kelly, C. P. & LaMont, J. T. Clostridium difficile–more difficult than ever. en. *N. Engl. J. Med.* **359,** 1932–1940 (Oct. 2008).

2. Rupnik, M., Wilcox, M. H. & Gerding, D. N. Clostridium difficile infection: new developments in epidemiology and pathogenesis. en. *Nat. Rev. Microbiol.* **7,** 526–536 (July 2009).

3. Buffie, C. G., Bucci, V., Stein, R. R., McKenney, P. T., Ling, L., Gobourne, A., *et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature [Internet]* 2015.

4. Cohen, S. H., Gerding, D. N., Johnson, S., Kelly, C. P., Loo, V. G., McDonald, L. C., Pepin, J. & Wilcox, M. H. Clinical practice guidelines for Clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of America (SHEA) and

the infectious diseases society of America (IDSA). *Infect. Control Hosp. Epidemiol.* **31,** 431–455 (2010).

5. Freddolino, P. L., Amini, S. & Tavazoie, S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. en. *J. Bacteriol.* **194,** 303–306 (Jan. 2012).

6. Collery, M. M., Kuehne, S. A., McBride, S. M., Kelly, M. L., Monot, M., Cockayne, A., Dupuy, B. & Minton, N. P. What's a SNP between friends: The influence of single nucleotide polymorphisms on virulence and phenotypes of Clostridium difficile strain 630 and derivatives. en. *Virulence* **8,** 767–781 (Aug. 2017).

7. Raškevičius, V., Mikalayeva, V., Antanavičiūtė, I., Ceslevičienė, I., Skeberdis, V. A., Kairys, V. & Bordel, S. Genome scale metabolic models as tools for drug design and personalized medicine. en. *PLoS One* **13,** e0190636 (Jan. 2018).

8. Zhang, C. & Hua, Q. Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. en. *Front. Physiol.* **6,** 413 (2015).

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

10. Larocque, M., Chénard, T. & Najmanovich, R. A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. en. *BMC Syst. Biol.* **8,** 117 (Oct. 2014).

11. Kashaf, S. S., Angione, C. & Lió, P. Making life difficult for Clostridium difficile: augmenting the pathogen's metabolic model with transcriptomic and codon usage data for better therapeutic target characterization. *BMC Syst. Biol.* **11,** 25 (Feb. 2017).

12. Bochner, B. R., Gadzinski, P. & Panomitros, E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. en. *Genome Res.* **11,** 1246–1255 (July 2001).

13. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. en. *Methods Mol. Biol.* **1151,** 165–188 (2014).

14. Nursimulu, N., Xu, L. L., Wasmuth, J. D., Krukov, I. & Parkinson, J. Improved enzyme annotation with EC-specific cutoffs using DETECT v2. en. *Bioinformatics* (May 2018).

15. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E. & Ye, J. Database resources of the National Center for Biotechnology Information. en. *Nucleic Acids Res.* **40,** D13–25 (Jan. 2012).

16. Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B. & Lercher, M. J. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. en. *PLoS Comput. Biol.* **13,** e1005494 (Apr. 2017).

17. Hartleb, D., Jarre, F. & Lercher, M. J. Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. en. *PLoS Comput. Biol.* **12,** e1005036 (Aug. 2016).

18. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? en. *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

19. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic Acids Res.* **44,** D515–22 (Jan. 2016).

20. Norsigian, C. J., Pusarla, N., McConn, J. L., Yurkovich, J. T., Dräger, A., Palsson, B. O. & King, Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. en. *Nucleic Acids Res.* (Nov. 2019).

21. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. *ssbio: A Python Framework for Structural Systems Biology* en. July 2017.

22. Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E. J., Bliven, S. E., Chen, K., Chang, R. L., Bourne, P. E. & Palsson, B. O. Systems biology of the structural proteome. en. *BMC Syst. Biol.* **10,** 26 (Mar. 2016).

23. Light, S. H., Minasov, G., Shuvalova, L., Duban, M.-E., Caffrey, M., Anderson, W. F. & Lavie, A. Insights into the mechanism of type I dehydroquinate dehydratases from structures of reaction intermediates. en. *J. Biol. Chem.* **286,** 3531–3539 (Feb. 2011).

24. Asojo, O. A., Nelson, S. K., Mootien, S., Lee, Y., Rezende, W. C., Hyman, D. A., Matsumoto, M. M., Reiling, S., Kelleher, A., Ledizet, M., Koski, R. A. & Anthony, K. G. Structural and biochemical analyses of alanine racemase from the multidrug-resistant Clostridium difficile strain 630. en. *Acta Crystallogr. D Biol. Crystallogr.* **70,** 1922–1933 (July 2014).

25. Demmer, J. K., Pal Chowdhury, N., Selmer, T., Ermler, U. & Buckel, W. The semiquinone swing in the bifurcating electron transferring flavoprotein/butyryl-CoA dehydrogenase complex from Clostridium difficile. en. *Nat. Commun.* **8,** 1577 (Nov. 2017).

26. Knauer, S. H., Buckel, W. & Dobbek, H. Structural basis for reductive radical formation and electron recycling in (R)-2-hydroxyisocaproyl-CoA dehydratase. en. *J. Am. Chem. Soc.* **133,** 4342–4347 (Mar. 2011).

27. Karasawa, T., Ikoma, S., Yamakawa, K. & Nakamura, S. A defined growth medium for Clostridium difficile. en. *Microbiology* **141 (Pt 2),** 371–375 (Feb. 1995).

28. Hafiz, S. & Oakley, C. L. Clostridium difficile: isolation and characteristics (Plate VIII). *J. Med. Microbiol.* **9,** 129–136 (1976).

29. Dembek, M., Barquist, L., Boinett, C. J., Cain, A. K., Mayho, M., Lawley, T. D., Fairweather, N. F. & Fagan, R. P. High-throughput analysis of gene essentiality and sporulation in Clostridium difficile. en. *MBio* **6,** e02383 (Feb. 2015).

30. Jenior, M. L., Leslie, J. L., Young, V. B. & Schloss, P. D. Clostridium difficile Colonizes Alternative Nutrient Niches during Infection across Distinct Murine Gut Microbiomes. en. *mSystems* **2** (July 2017).

31. Scaria, J., Chen, J.-W., Useh, N., He, H., McDonough, S. P., Mao, C., Sobral, B. & Chang, Y.-F. Comparative nutritional and chemical phenome of Clostridium difficile isolates determined using phenotype microarrays. en. *Int. J. Infect. Dis.* **27,** 20–25 (Oct. 2014).

32. Theriot, C. M., Koenigsknecht, M. J., Carlson, P. E., Hatton, G. E., Nelson, A. M., Li, B., Huffnagle, G. B., Li, J. Z. & Young, V. B. *Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection* 2014.

33. Duthie, G. G. & Wood, A. D. Natural salicylates: foods, functions and disease prevention. en. *Food Funct.* **2,** 515–520 (Sept. 2011).

34. Sonowal, R., Nandimath, K., Kulkarni, S. S., Koushika, S. P., Nanjundiah, V. & Mahadevan, S. Hydrolysis of aromatic β-glucosides by non-pathogenic bacteria confers a chemical weapon against predators. en. *Proceedings of the Royal Society B: Biological Sciences* (July 2013).

35. Reizer, J., Ramseier, T. M., Reizer, A., Charbit, A. & Saier Jr, M. H. Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in Escherichia coli. en. *Microbiology* **142 (Pt 2),** 231–250 (Feb. 1996).

36. Kaval, K. G. & Garsin, D. A. Ethanolamine Utilization in Bacteria. en. *MBio* **9** (Feb. 2018).

37. Nawrocki, K. L., Wetzel, D., Jones, J. B., Woods, E. C. & McBride, S. M. Ethanolamine is a valuable nutrient source that impacts Clostridium difficile pathogenesis. en. *Environ. Microbiol.* **20,** 1419–1435 (Apr. 2018).

38. Tonner, P. D., Darnell, C. L., Engelhardt, B. E. & Schmid, A. K. Detecting differential growth of microbial populations with Gaussian process regression. en. *Genome Res.* **27,** 320–333 (Feb. 2017).

39. Klauer, A. A. & van Hoof, A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. en. *Wiley Interdiscip. Rev. RNA* **3,** 649–660 (Sept. 2012).

40. Nam, H., Lewis, N. E., Lerman, J. A., Lee, D.-H., Chang, R. L., Kim, D. & Palsson, B. O. Network context and selection in the evolution to enzyme specificity. en. *Science* **337,** 1101–1104 (Aug. 2012).

41. Collins, J., Robinson, C., Danhof, H., Knetsch, C. W., van Leeuwen, H. C., Lawley, T. D., Auchtung, J. M. & Britton, R. A. Dietary trehalose enhances virulence of epidemic Clostridium difficile. en. *Nature* **553,** 291–294 (Jan. 2018).

42. Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., Teiling, C., Steffy, B., Taylor, M., Schwartz, A., Richardson, T., White, C., Baele, G., Maere, S. & Verstrepen, K. J. Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. en. *Cell* **166,** 1397–1410.e16 (Sept. 2016).

43. Kay Holtman, C., Thurlkill, R. & Pettigrew, D. W. Unexpected Presence of Defective glpRAlleles in Various Strains of Escherichia coli. en. *J. Bacteriol.* **183,** 1459–1461 (Feb. 2001).

44. Kumar, N., Browne, H. P., Viciani, E., Forster, S. C., Clare, S., Harcourt, K., Stares, M. D., Dougan, G., Fairley, D. J., Roberts, P., Pirmohamed, M., Clokie, M. R. J., Jensen, M. B. F., Hargreaves, K. R., Ip, M., Wieler, L. H., Seyboldt, C., Norén, T., Riley, T. V., Kuijper, E. J., Wren, B. W. & Lawley, T. D. Adaptation of host transmission cycle during Clostridium difficile speciation. en. *Nat. Genet.* **51,** 1315–1320 (Sept. 2019).

45. Fletcher, J. R., Erwin, S., Lanzas, C. & Theriot, C. M. Shifts in the Gut Metabolome and Clostridium difficile Transcriptome throughout Colonization and Infection in a Mouse Model. en. *mSphere* **3** (Mar. 2018).

46. 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).

47. Seemann, T. *mlst*

48. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. en. *Wellcome Open Res* **3,** 124 (Sept. 2018).

Chapter 5

A workflow for generating multi-strain genome-scale metabolic models of prokaryotes

5.1 Abstract

Genome-scale models (GEMs) of bacterial strains' metabolism have been formulated and used over the past 20 years. Recently, with the number of genome sequences exponentially increasing, multi-strain GEMs have proved valuable to define the properties of a species. Here, through four major stages, we extend the original Protocol used to generate a GEM for a single strain to enable multi-strain GEMs: (i) obtain or generate a high-quality model of a reference strain; (ii) compare the genome sequence between a reference strain and target strains to generate a homology matrix; (iii) generate draft strain-specific models from the homology matrix; and (iv)

manually curate draft models. These multi-strain GEMs can be used to study pan-metabolic capabilities and strain-specific differences across a species, thus providing insights into its range of lifestyles. Unlike the original Protocol, this procedure is scalable and can be partly automated with the Supplementary Jupyter notebook Tutorial (See `https://www.nature.com/articles/s41596-019-0254-3`). This Protocol Extension joins the ranks of other comparable methods for generating models such as CarveMe and KBase. This extension of the original Protocol takes on the order of weeks to multiple months to complete depending on the availability of a suitable reference model.This protocol is an extension to: Nat. Protoc. doi: `https://doi.org/10.1038/nprot.2009.203`

5.2 Introduction

In recent years, the exponential increase in the number of genome sequences has enabled us to investigate the variability across strains within the same species. As more genome sequences become available, significant differences in genomic content and functions across strains have been identified [1]. Therefore, researchers started to explore strain-specific variations using approaches such as pan-genome analyses [2]. These analyses showed that some species have a vast diversity of genes among its strains, resulting in remarkably different divergent phenotypes across strains [3]. However, despite the utility of pan-genome analysis based on gene lists, it does not provide mechanistic insight into phenotypic potential based on genetic and genomic variability within a species.

Over the past decade, genome-scale models (GEMs) of metabolism have proven to be valuable in understanding mechanistic links between genotype and phenotype [4]. GEMs are mathematical models of metabolic network reconstructions [5]. They allow computation of the

systems-level metabolic functions from genome sequences and extend the power of pan-genome analyses towards sequence-based evaluation of the phenotypic variation of a species. So far, the majority of studies based on metabolic network reconstructions, and GEMs derived from them, have been focused on a single strain of a species. This includes a large number of studies based on our previously published metabolic network reconstruction Protocol [6].

A strain-specific GEM can be expanded into models for multiple strains of the same species. Rapid mapping of the gene content in a GEM from a reference strain onto multiple strains' genome sequences of interest is now possible. This process allows one to utilize highly curated knowledge bases assembled over many decades, upon which a metabolic reconstruction is based, to quickly study a freshly sequenced isolate. Using this process, recent studies have successfully identified strain-specific metabolic differences and their association with lifestyle of the strains for multiple species [7–12]. These studies lead to an understanding of strain diversity, for species with both large and small pangenomes. Using GEMs to characterize pan-genomes is thus likely to be a widely used method as thousands of strain sequences will become available for species across the microbial phylogenetic tree.

It is worth noting that other methods for the generation of GEMs from existing reconstructions are available, namely CarveMe [13] and functions within KBase[14]. CarveMe relies on the use of a universal model that is then filtered to a specific model by solving a mixed integer linear program. KBase executes a proteome comparison and utilizes that information to infer reactions to keep within a new model. The reliance on the universal model within CarveMe may limit the achievable specificity particularly in regard to biomass equations. CarveMe also possesses the unique functionality to produce ensemble models and microbial community models. KBase benefits from ease-of-use and potential integration with other KBase functions, however

the implementation is restricted to the KBase interface and limited in customizability.

In this Protocol, we extend our original metabolic reconstruction Protocol [6] to instruct users on building multi-strain GEMs from an existing reference model. We will provide guidance for the reconstruction and application of strain-specific models and show how a reference strain is mapped to other strains within the same species. Furthermore, we provide a detailed tutorial (Supplementary Tutorial) along with the step-by-step instructions to guide readers through the Protocol and its efficient implementation. The application of the workflow is rapid, and it can be partly automated.

5.3 Applications

A highly curated reference reconstruction represents a highly organized and structured assembly of organism information. This accumulated knowledge can be efficiently extended to generate strain-specific models by combining comparative genomics and genome-scale metabolic modeling (Figure 5.1). By analyzing multiple strains, it becomes possible to investigate the range of evolutionary outcomes for a species. GEMs allow for the prediction of growth capabilities and auxotrophies across a bacterial species. These predictions have provided insight into the lifestyle and diversity of the members of a species. For example, metabolic capabilities predicted using multi-strain GEMs have been used to build classification schema capable of organizing strains into nutrient niche [7], serovars [9], and pathogenicity [12]. Multi-strain GEMs provide a platform with which to begin to combat limitations identified with reconstruction efforts [15] regarding completeness and the coverage of the reactome.

Another inherent strength of multi-strain reconstruction is scalability. The number of strains considered may be increased with ease. Scalability, in turn, enables new applications. On

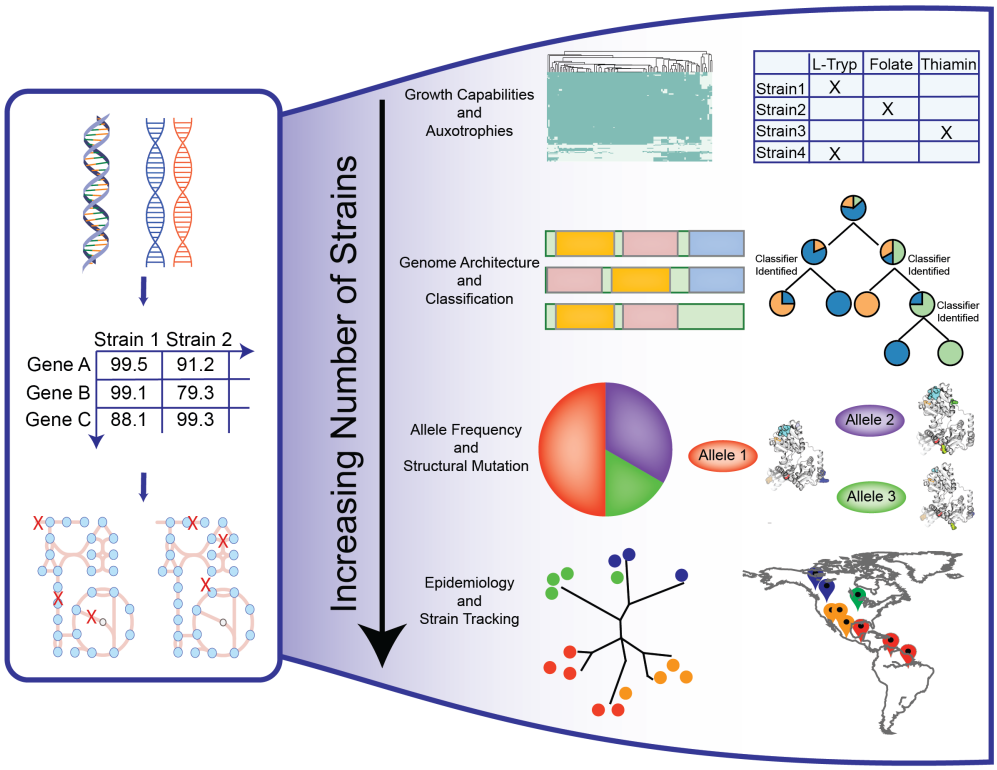

Figure 5.1: Applications of multi-strain GEMs. The workflow of genome comparison to generate a homology matrix of PIDs, which is in turn used to generate strain-specific models of the target strains. The values in the homology matrix are percentages. The number of strains considered in this manner enables various types of analyses including: (i) comparison of strain nutrient utilization and identification of strain-specific auxotrophies, (ii) interrogation of genome architecture and classification of strains by niche or by pathotype, (iii) investigation of allele frequencies among strains and mapping to protein structural information, and (iv) linking to epidemiology and tracking of strains or infections.

the order of hundreds of strains, it becomes possible to use multi-strain GEMs to investigate allele frequencies of genes within a network context [16]. The reconstructed networks provide insight into potential evolutionary hotspots that become linked to calculated phenotypes through the use of GEMs [16]. Additionally, the higher number of strains considered allows for applications with a wider perspective. For example, studying the global epidemiology of infection and the tracking of strains by their indicated abilities and classifications become possible. It is worth noting that the number of strains considered may also potentially influence the complexity and time

for downstream analysis. Preliminary results become rapidly available through this approach, however if additional strains are candidates for extensive curation this increases the time required for future analysis.

5.4 Advantages and Limitations

Multi-strain GEMs provide us with a comprehensive and high-resolution knowledge base of metabolic diversity across strains of a species of interest. The models enable accurate and rapid computational prediction of auxotrophies and nutrient utilization capability across strains from only genome sequences without the need for experiments. The results then allow us to calculate correlations between strain-specific metabolic variations and attributes of the strain's lifestyle (such as host specificity) or health outcomes such as strain-specific implications in inflammatory bowel disease [7, 8, 12, 17]. The reconstruction of multi-strain GEMs is much faster than reconstructing a reference model from scratch, yet still highly informative.

However, the user should also keep in mind the limitations before starting the multi-strain GEM reconstruction. First, it can be time-consuming to build multi-strain GEMs for species lacking a reference model, as approximately six months to a year is needed to build a GEM de novo. Second, this Protocol Extension works best with well-annotated species, since a lack of information may result in an incomplete model and inaccurate predictions. Nevertheless, strain-specific GEMs will also enable the discovery of knowledge gaps for less well-studied species. Third, multi-strain GEMs will be most valuable for species with significant differences in genomic content across strains. If strains within the species have limited genetic variability, the strain-specific GEM will be very similar and provide limited new information. Such similarity can be quickly evaluated by examining the openness of the pan-genome for the strains of interest.

Finally, basic coding skills are required for this Protocol Extension. Previous experience with bioinformatics analysis, coding languages (especially python), and usage of GEMs will accelerate the process significantly.

5.5 Experimental Design

This Protocol Extension consists of four major stages to utilize the output of a high-quality genome-scale metabolic reconstruction [6] to create multiple strain-specific models derived from the reference organism (Figure 5.2). These stages are described further in the following sections. These stages are also summarized within a pseudocode format. Following the steps delineated here will result in draft strain-specific models based on genetic similarity to the original strain that can be used as a starting point to feed directly into Stage 2 of the original Protocol [6] for further refinement and evaluation, or for immediate comparative investigation. The time-consuming nature of the base reconstruction approach of the original Protocol [6] results in limited scalability; this approach of generating models for multiple strains through homology relationships represents a means of more rapidly extrapolating the knowledge contained within the highly-curated reconstruction. One caveat to consider when applying this approach is the metabolic diversity inherent to the species of interest. If the species is not particularly genetically diverse, then the resulting models will likewise be highly similar.

Along with the step-by-step procedures in this Protocol Extension, we also provide a tutorial (Supplementary Tutorial) to generate strain-specific models for five *E. coli* strains from a reference model. The Supplementary Tutorial includes 3 jupyter notebooks that are focused on stages 2 and 3 (the genome sequence comparison and generation of homology matrix stage, and the creation of strain-specific draft models) to guide the steps that could be automated in

this Protocol Extension.

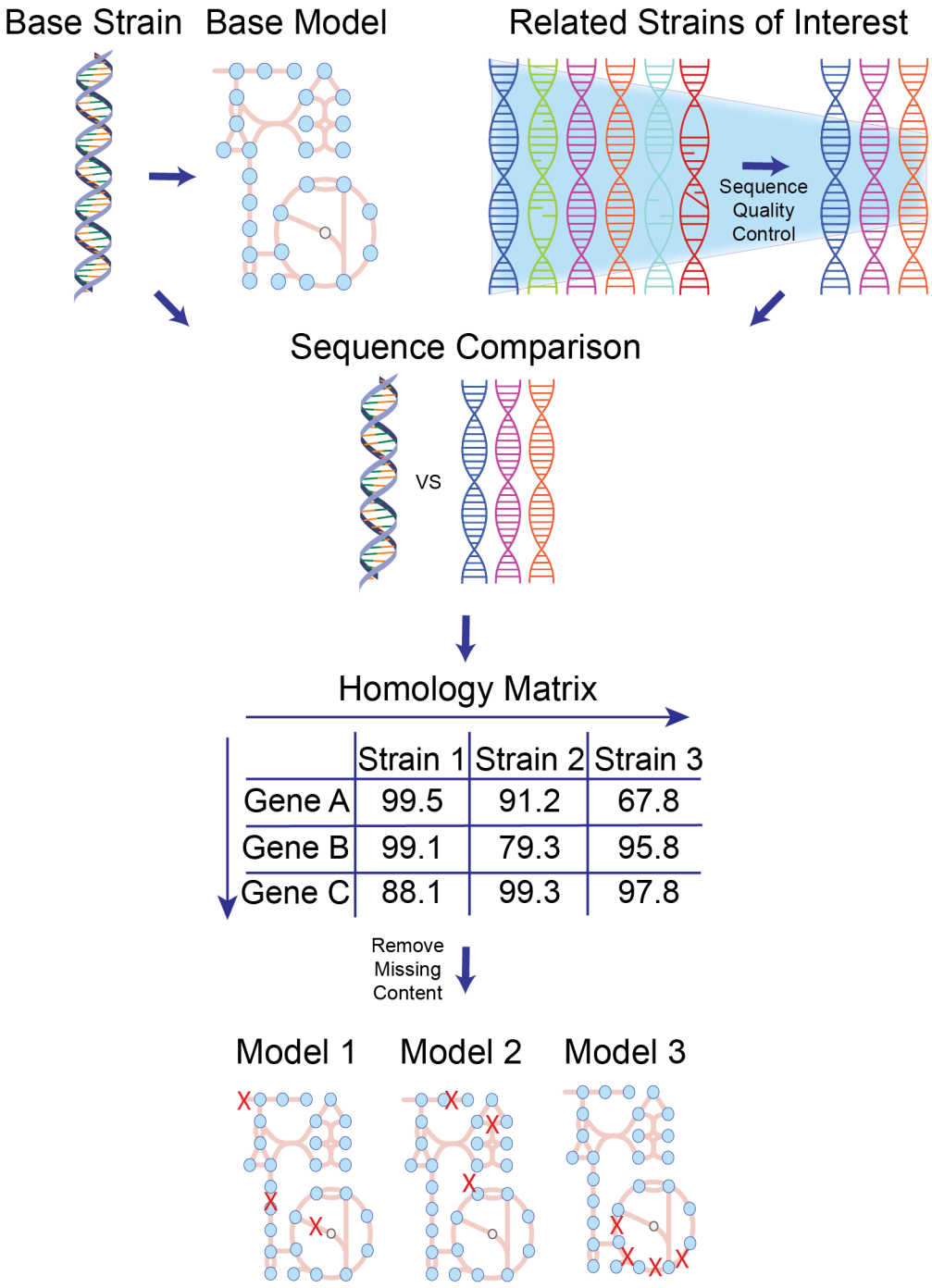

Figure 5.2: Overall workflow for multi-strain GEM generation

5.6 Overview of the Procedure

5.6.1 Stage 1: Steps 1-4, obtain a high-quality starting reference reconstruction

To generate strain-specific reconstructions, a high-quality single-strain base reconstruction generated through the use of the original Protocol [6] is a necessary starting point. Published reconstruction efforts usually include this output as a supplementary data file in either SBML or JSON file formats. Additionally, a number of reconstruction repositories exist, such as BiGG, BioModels, and MetaNetX [18–20]. If a reconstruction for a reference strain in the species of interest is not available, then the original Protocol can be executed to produce one [6]. The resulting output can then be used as the starting point to generate multi-strain models. It is possible that for certain organisms there could be multiple available models that have been independently reconstructed. This represents a potential opportunity to broaden the reference knowledgebase. In this case the user can either reconcile the base reconstructions for a single strain into a single reconstruction of highest confidence through careful manual curation of the content or run this Protocol Extension using each base reconstruction in turn and compare the resulting draft models of interest. After obtaining (or generating) a reference reconstruction, it is necessary to evaluate its quality to determine its suitability for use as a reference reconstruction. To evaluate the reference reconstruction, refer to Stage 4 of the original Protocol [6]. Recently, a testing suite called Memote has become available that evaluates a number of quality control/quality assurance features of a GEM in a drag and drop fashion [21]. Once a curated, quality reference reconstruction is either obtained or generated, it can be used in the following steps to generate strain-specific GEMs.

Box 1: A commentary on genome annotation and assembly

Genome annotation and assembly are both well documented and established techniques within the bioinformatics field[22] If the research effort is using publicly available genomes, most will likely be annotated. However, when utilizing newly sequenced genomes or those lacking annotation, it is necessary to perform annotation. While a plethora of tools exist for executing genome annotation [23, 24], it is important to use a consistent tool to prevent potential errors/bias. One potentially useful annotation software package is Prokka [25]. If one is interested in following this Protocol Extension to generate models of newly sequenced strains it will also be necessary to perform genome assembly. This raises the question of the sequence quality required to generate multi-strain models. One means of assessing quality is through coverage. While the specific requirement may vary from species to species, we analyzed how varying coverage impacts the resulting assembly metrics of N50 and number of contigs (Supplementary Figure D.2). For the purposes of using an assembled genome, it is important to have sufficient coverage that demonstrates saturation in these metrics. In the case of the *E. coli* strain discussed in the supplementary figure D.2 we see this to be at around 70X coverage.

5.6.2 Stage 2: Steps 5–13, genome sequence comparison and generation of homology matrix

Stage 2 is to identify and acquire the sequenced genomes of different strains from the species of interest. Publicly available genome data is available in sources such as NCBI or PATRIC [26, 27]. How many and which strains to include depends on the given research question posed. Criteria for genome selection could possibly include particular isolation location, existence of associated metadata, and phenotype or pathotype information. One should keep in mind the

phylogenetic distance between reference strain and target strains as this will directly impact the utility of mapping the content of the original reconstruction. As a means of quality assurance, it is important to keep track of the identifiers of the publicly available genomes used. Within Notebook 1 (Supplementary Tutorial) we begin by acquiring a small set of *E. coli* genomes from NCBI. In the described workflow and corresponding tutorials, we assume that the user is starting with annotated GenBank files for the strains of interest (see Box 1).

After identifying and obtaining the genomes for the target strains of interest, the next step is to identify the orthologous genes between each strain and the reference strain. This step is detailed within Notebook 1 (Supplementary Tutorial). While a plethora of techniques exist to perform this function, we recommend utilizing NCBI protein BLAST to identify bidirectional best hits as it is widely adopted by the community, scriptable, and reliable. This method is utilized within the provided scripts (Supplementary Tutorial). Following the identification of homologous genes in each of the target strains, the results can be unified into a single Pandas dataframe of the percentage identity values (PID). This dataframe is then filtered down to contain all the genes within the reference reconstruction. The output of these steps is the homology matrix consisting of N x M PIDs, where there are N rows of the genes within the reference reconstruction and M columns of the target strains (Figure 5.2). The penultimate step is to apply a threshold to binarize this matrix into a presence/absence matrix detailing which genes are absent within the target strain. We suggest utilizing a cutoff of 80% percentage sequence identity covering at least 25% of the query gene length or above to consider the gene present within the target strain. However, this threshold is an adjustable parameter and the effect of genes retained in draft strain-specific models is dependent on how genetically similar the target strains are to the reference strain (Supplementary Figure D.1).

A supplementary final step is to execute a nucleotide BLAST. Many reference genomes have undergone extensive manual curation within the annotation, so there may be discrepancies with automatically annotated target strains. By executing a BLAST on raw nucleotide sequences there is a secondary comparison made to catch potentially unannotated open reading frames within a given target strain. In addition, for each open reading frame (ORF) identified to pass the nucleotide sequence similarity threshold but missing from the annotations, a quality check for premature stop codons within the sequence is performed as these ORFs likely result in a nonfunctional protein. This process is also detailed within Notebook 1 (Supplementary Tutorial). The nucleotide BLAST provides an added catch to avoid excluding genes from strain-specific models due to lack of annotation. The final binarized homology matrix can then be used in concert with COBRA methods [28–31] to create and save strain-specific models of the target strains.

5.6.3 Stage 3: Steps 14–23, creation of strain-specific draft models

The genome comparison executed in Stage 2 provides information on which genes within the base reconstruction are lacking a homologous gene within each target strain genome. By utilizing the "remove genes" function from the "cobra.manipulation.delete" module of COBRApy, the appropriate genes can be removed from a model. Notebook 2 (Supplementary Tutorial) demonstrates how to properly implement this technique. For every target strain of interest, a copy of the base reconstruction is created and appropriate genes, as per the homology matrix, are deleted from each model, creating a draft strain-specific model. This process is repeated for each strain of interest. Additionally, the genes retained in each strain-specific model are updated at this stage to reflect the locus tags in the target strain's annotation. This process is executed

using the "rename genes" function from the "cobra.manipulation.modify" module according to another generated matrix of all the gene names, mapping the gene identifiers constructed within Stage 2. Depending on the annotation platform it may be worthwhile to add additional locus tag information to stratify multiple namespaces. For example, if the genomes used were re-annotated with Prokka it could be useful to add NCBI locus tags to the gene objects within the model. Additional information can be stored within the "notes" field of a gene object. The updated draft models are then ready for further evaluation.

The next step is functional evaluation of the draft strain-specific models and this begins by determining which of them are able to be optimized through linear programming for biomass objective flux, i.e., *in silico* growth. At this point, a combination of automated gap-filling methods and manual curation are used to determine which nutrients need to be supplemented to the *in silico* media to achieve positive biomass yield. Gap-filling methods have been well documented [32–35], and the results generated can be used to enable growth in strain-specific models found to have auxotrophies.

This step is executed in an iterative fashion across all target strains and reflects a critical step in any reconstruction effort. It is important to keep in mind the differences between the two model types to be gap-filled: 1) models of true auxotrophs that require only a supplementation of extracellular nutrient to enable biomass production and 2) models in which metabolic reaction gap-filling is necessary and thus offers a potential for discovery of alternative pathways. Ideally, gap-filling should always be supported by literature information and/or validated experimentally. In this context, we refer to the gap-filling required to obtain a functional network that can produce biomass. It is also worth noting that in some cases where there are known biomass composition variants, instead of gap-filling the model to enable growth, the biomass reactions

should be modified. Alternate biomass formulations may substantially affect model predictions and have been shown to be variable across species and conditions [36].The base biomass reaction may also be highly variable across strains in certain species. For example, O antigen structures are highly variable across Gram-negative strains and the corresponding biosynthetic pathways vary extensively, requiring a separate pan-species reconstruction effort.[9] Therefore, instead of directly taking the biomass reaction from the base strain, we recommend that the users customize biomass reaction for strains of interest by generating or collecting strain-specific experimental data, when available. A recently developed workflow can also help users generate the biomass reactions in a data-driven and unbiased fashion.

5.6.4 Stage 4: Steps 24–28, curation of strain-specific models

At this juncture, a group of functional models for the identified target strains has been produced, and may be used in their current form to generate preliminary predictions and direct future studies. Any known strain capabilities present an opportunity to perform a validation step to inspect whether the strain-specific models can still accurately predict known phenotypes. Additionally, all, or select models depending on interest and/or time constraints, can now be extensively manually curated as per the original Protocol [6] to produce a high-quality reconstruction. In this case, the models produced would be used as input to the original Protocol [6] at Stage 2: Reconstruction Refinement. This would refine these models from derivative draft strain-specific models to curated reconstructions of specific strains. This effort will involve adding strain-specific metabolism not present in the original reconstruction. One useful technique here would be to annotate the pangenome to potentially catch genes with divergent nucleotide sequence but similar functional machinery which may have appeared due to horizontal gene

transfer events. While additional manual curation of the generated strain-specific models would yield more accurate predictions, it is worth noting that the group of draft models represents a valuable resource.

Various analyses can be conducted such as determining differing growth capabilities across nutrient environments. An example of this for carbon source utilization is demonstrated in Notebook 3 (Supplementary Tutorial). In this analysis, growth in different nutrient conditions can quickly be predicted. Starting from a minimal media condition, the current growth-supporting nutrients for carbon, nitrogen, phosphorous, or sulfur can be removed, and an appropriate list of nutrients looped through to determine whether alternative sources of carbon, nitrogen, phosphorus, and sulfur support growth. This process is repeated for each strain in the group of strain-specific models. Experimental validation of the multi-strain predictions is ideal. The resulting *in silico* predicted growth capabilities can then be used to examine which strains are similar in terms of metabolic phenotype. This approach has proven fruitful in providing an additional level of discrimination in numerous past studies and represents one of the immediate benefits of extending a reconstruction to construct strain-specific models.

5.6.5 Stage 5: applications of multi-strain GEMs

Once a collection of functional models of the identified target strains has been generated, they can be used in a variety of ways (see Applications). This fifth stage includes a range of techniques to select from, determined by the research to be conducted. Given the breadth of the potential applications, they are not addressed in this Protocol Extension.

5.7 Materials

5.7.1 Annotated genome sequences of interest

Annotated sequences of interest can either be downloaded from public databases or generated by the user through sequencing. In this Protocol Extension, we start with annotated GenBank files that contain the annotation and sequence sections that can be directly downloaded from NCBI. Several other guides document how to assemble and annotate genomes of interest [37, 38].

5.7.2 Reference GEMs

Reference GEMs have already been reconstructed for many well studied organisms. The available GEMs can mostly be found and downloaded from publications or public databases such as BiGG Models [18]. Reference models can be in various formats such as SBML, MAT and JSON. If the Reference model has not been built for the species of interest, please refer to the original Protocol[6] for details of building a detailed, reference reconstruction.

5.7.3 Equipment and Software

Standard personal computer with the following software/packages properly installed:BLAST (v 2.9.0 tested), Python (v 3.5.2 tested). Python Packages: pandas (v0.23.0 tested), seaborn (v0.8.1 tested), biopython (1.71 tested), jupyter notebook (v5.2.3 tested). All python packages can be installed directly with pip command. If the users are more comfortable with anaconda, all packages are available in anaconda installation as well. CobraPy: the installation steps and tutorial can be found on `https://cobrapy.readthedocs.io/en/latest/`. To ensure the performance of the scripts in the Supplementary Tutorial, use version 0.13.0

5.8 Procedure

5.8.1 Stage 1: reconstruction of base model: Timing 6 months to 1 year

1. *Obtain reference model.* Download a reference model from BiGG Models (http://bigg.ucsd.edu/), publications or other databases. The resulting draft strain-specific models will reflect the namespace of the base reconstruction. **CRITICAL STEP:** Models in the BiGG database [18] have been pre-checked for quality, so it is a recommended resource if your organism of interest is available. While BiGG is recommended, any consistent reconstruction where the gene product rules are linked to a genome annotation, producing a model that can be loaded to COBRApy will work within this Protocol Extension.

2. *Build reference model if not available.* If the reference model is not available, reconstruct a model from scratch following the original reconstruction Protocol [6] or start from draft models reconstructed in previous studies [39, 40] and follow the original Protocol[6].

3. *Quality control.* Regardless of the source, perform quality control analysis on the base model by uploading the model to Memote (https://memote.io/) [21] for quality checking. Once the report is available, check the following two important measures: 1. All metrics in the consistency section 2. Uniform Metabolite Identifier Namespace. These metrics ensure that the model is properly standardized. In addition, check if the model is functional by performing growth simulations to ensure firstly that there is no growth when exchange reactions are closed. Refer to the computational method developed by Fritzmeier et al. [41] to identify and remove erroneous energy-generating cycles. And secondly that the growth prediction is consistent with experimental observations including nutrient utilization and metabolite secretion (if data available). **CRITICAL STEP:** The quality of the multi-

strain models generated from this Protocol Extension will be highly dependent on the reference model. So, it is especially important to start with a high-quality reconstruction and experimentally validated model.

4. *Obtain base strain genome annotation.* Download the reference strain genome annotation.Retrieve the GenBank file that contains the genome sequence and annotation, which were originally used to reconstruct the reference GEM and extract the modeled coding DNA sequences and corresponding unique locus tags. **CRITICAL STEP:** This is important because the creation of draft multi-strain model is dependent on the sequence annotation of the base strain.

5.8.2 Stage 2: sequence comparison and generation of homology matrix: Timing days to weeks

5. *Download annotated genomes for different strains of interest in GenBank format.* Genomes of interest can be downloaded from various public databases such as National Center for Biotechnology Information (NCBI) and Pathosystems Resource Integration Center (PATRIC) [26]. Instructions to download annotated genome sequences from NCBI can be found here: `https://www.ncbi.nlm.nih.gov/guide/howto/dwn-genome/`, and instructions to download genome sequences from PATRIC can be found here: `https://docs.patricbrc.org/user_guides/data/index.html#download-data`. Or users can follow the Supplementary Tutorial to download the GenBank files using jupyter notebooks. !CAUTION: We recommend downloading GenBank files that contain both sequence and annotation information. If annotation is not available for the target strains, see Box 1 for our recommendations and tips on genome annotation. To ensure consistency, the annota-

tion pipeline used for target strains should be the same as the pipeline used for reference strain.

6. *Quality control of the genome sequences.* Calculate and check the coverage (if available), N50 score and number of contigs of the genome sequences. To determine the threshold for the above quality metrics, consider performing similar analysis shown in Supplementary Figure D.1. Discard genome sequences that do not pass the quality test. **CRITICAL STEP:** More reliable results can be obtained from genome sequences with coverage > 70x. Adjust the threshold for quality metrics such as N50 score and number of contigs based on your organism of interest, as they are highly dependent on the organism. If time permits, use sensitivity analysis [42] to find the most appropriate threshold.

7. *Generate Fasta files from GenBank files.* Use the Genbank files to generate fasta files for both protein and nucleotide sequences (see Notebook 1 in Supplementary Tutorial). Protein fasta files are then used as input for the following BLAST operation in Step 9 to identify homologous proteins across strains.

8. *Identify candidate metabolic functions.* The previous genome annotation (Step 4) should provide E.C. numbers for genes involved in metabolic function. Extract genes with E.C. numbers from annotations and the following steps are focused on these metabolic genes only.

9. *BLAST the genomes of interest against the reference strain.* Perform bidirectional protein BLAST [43] to identify the sequence similarity of metabolic proteins in strains of interest compared to the reference strain. Use BLASTp (output format 6) to record both query/subject ID and percentage identity matches (PID). **CRITICAL STEP:** Bidirec-

tional BLAST uses both the reference strain or the other strain as reference BLAST database and selects the best bidirectional hits (BBH) based on BLAST result in both directions to identify orthologs. Note that we recommend filtering mapping results based on coverage of alignment length. (see Notebook1 in Supplementary Tutorial)

10. *Filter the BLAST result for only proteins in the base model.* Identify the list of proteins included in the base model and keep only the BLAST results for these protein genes for the following analysis.

11. *Create a homology matrix summarizing the results for all strains of interest.* Identify the BBHs of all proteins between reference strain and strains of interest. Compile the PID of all BBHs in the base model for all strains into a homology matrix, where the columns represent the strains, and the rows represent the protein.

12. *Create binarized homology matrix for genes in the model.* Select a threshold for PID to determine the presence/absence of proteins in all strains. The matrix is binary with 1 representing presence, and 0 representing absence. Similar to the homology matrix, it should have M strains * N proteins. **CRITICAL STEP:** Adjust the threshold for PID accordingly depending on your data and purpose (see Supplementary D.1 for how PID threshold affects the number of genes retained in strain-specific models). The threshold of 80% used in the Supplementary Tutorial is quite stringent as some tools use the sequence identity cutoff of 50% to identify gene orthologs [44].

13. *Nucleotide BLAST to check unannotated open reading frames.* To ensure that we do not miss any genes in the target strains due to lack of annotation during BLASTp, we perform nucleotide BLAST between the reference strain and nucleotide sequences of the target

strains (fna files containing contigs). In addition, we also look for premature stop codons in genes of interest to exclude non-functional proteins. Record any inconsistencies observed in gene absence/present results generated by BLASTp and BLASTn, as they are potential candidates for manual curation.

5.8.3 Stage 3: creation of draft multi-strain models: Timing days to weeks

14. *Identify missing reactions.* Based on the presence/absence matrix from Step 12 and the gene-protein-reaction (GPR) established in the reference strain reconstruction, identify the genes missing in each strain and reactions encoded by the missing genes.

15. *Remove missing genes/reactions.* For each strain of interest, start with the reference strain. Remove the identified missing gene/reaction for each strain from the starting base strain using COBRApy function "remove genes". Save the modified model as the draft strain-specific model.!CAUTION: Multiple functions in COBRApy allow the user to delete reactions, but make sure to use function "remove genes" with the the parameter "remove reactions=True" to remove both the missing genes and reactions.

16. *Update the GPR in the draft models.* Using the query/subject ID obtained in step 9, match the genes in the base model with genes in the strains of interest to update the gene names in the strain-specific model. Optionally, to ensure that all possible encoding genes of a metabolic reaction are included in strain-specific models, one can refer to the full BLAST result from Step 9 and identify cases of additional pertinent homologs (potential paralogs) that are not BBH but also pass the PID threshold, and update the GPR accordingly (e.g., "Gene A" to "Gene A or Gene B").

17. *Check biomass reaction.* Make sure that the metabolites are general to all strains of interest. Remove the metabolites which are specific to the reference strain or its unique microenvironment. If strain-specific experimental omics data-sets are available, coefficients of metabolites in the biomass reaction could also be adjusted accordingly using the BOFdat workflow [45].

18. *Simulate growth.* For each strain-specific model, simulate for growth under the same medium condition as the base model using COBRApy function model.optimize(). A minimal medium condition is preferred if the recipe is available to identify potential auxotrophies. To modify the medium composition, change the constraint on the exchange reactions (see original Protocol Step 37 for more details [6]). If the simulated growth rate is less than 0.001 and the objective status is "optimal", skip Steps 19 to 23 and proceed to stage 4 directly. Otherwise continue with Step 19. ?Troubleshooting !CAUTION: Adjust the lower bound of the exchange reactions to allow uptake of extracellular nutrients. Ensure the exchange reactions of the metabolites missing from the medium are closed (lower bound set to 0).

19. *Identify strain-specific auxotrophies.* Simulate biomass yield in a rich medium (set all nutrient exchanges to -5 mmol/gDW/hr). If the yield obtained is less than 0.001 go to Step 22. Otherwise, find the minimal number of nutrient supplementations needed to support *in silico* growth using the "find nutrient supplementation" function. Review the literature for reports of an experimentally validated auxotrophic phenotype for your strain. If possible, acquire the strain and validate the auxotrophic phenotype experimentally.

20. *Check and report the genetic basis for the auxotrophy.* Retrieve the missing genes identified

by "find nutrient supplementation" as the genetic basis for the nutrient requirement and run BLASTn as a final quality check to ensure that no matches are found. If no matches are found, supplement the *in silico* medium for that strain-specific model with one of the sets of nutrients returned by "find nutrient supplementation". If the genes are found, add the reactions back into the strain-specific model, and adjust of the PID threshold used in Step 12 if needed. If positive yield is achieved skip Steps 21-23 and proceed to stage 4.

21. *Check biomass metabolite synthesis.* For strain-specific models which cannot simulate nonzero positive yield, simulate the production of each metabolite in the biomass reaction. To do so, create demand reactions which consume metabolites included in the biomass reaction. A demand reaction is a pseudo-reaction with a lower bound of 0 and upper bound of 1000 which allows for a metabolite to leave the cell. Instead of the biomass reaction, iteratively set one demand reaction as the objective to optimize for the production of each biomass precursor. If the flux through the demand reaction is less than 0.0001, model simulations suggest that this biomass precursor cannot be produced. ?Troubleshooting

22. *Identify missing essential reaction using gap filling.* Use the gapfill function in COBRApy to identify the minimum number of reactions that need to be added to the strain-specific model to enable the production of those biomass precursors which cannot be synthesized. Use the original reference model as the reaction repository to draw reactions from in the gap-filling step. Once the genetic basis for the simulated phenotype is identified, the curator should decide whether to exclude the precursor from the biomass reaction or add the gap filling reactions back. ?Troubleshooting

23. *Identify genetic evidence for missing essential reactions.* For the reactions identified in

the previous step, look for evidence in the genome and identify why they were deleted in the previous steps. Adjust sequence similarity threshold if needed and repeat the analysis from Step 12. If no genetic evidence is found, proceed with stage 4 to identify potential strain-specific alternative pathways.

5.8.4 Stage 4: curation of strain-specific models: Timing days to weeks

24. *Identify strain-specific genes absent from reference model.* Inspect the genes with E.C. number from strains of interest that are not present in the reference strain. Cross referencing models of related organisms may be helpful in this step.

25. *Identify novel metabolic reactions.* Identify metabolic reactions corresponding to the strain-specific genes identified in Step 23 using public databases including Uniprot (`https://www.uniprot.org/`), ModelSEED (`http://modelseed.org/`), KEGG (`https://www.genome.jp/kegg/`) and BIOCYC (`https://biocyc.org/`). Add the metabolic reaction to the model using COBRApy (see details in original Protocol [6] Steps 6-11). If the reaction is already present in the model, update the GPR of the reaction to include the strain-specific gene. !CAUTION: Make sure that the metabolite naming scheme for the novel reactions is consistent with the model standard to enable flux simulation through the newly-added reaction.

26. *Repeat growth simulation.* Ensure that draft models which were originally are able to simulate growth can still do so. Check if the models which failed to grow before can now simulate growth with newly-added reactions. If not, add back the missing essential reaction to enable follow-up analysis as it may be due to unknown alternative pathways. Ensure that the model does not have futile cycles after adding new reactions. **CRITICAL**

STEP: Growth simulation results could have been altered after adding novel strain-specific reactions. So even if the model was predicted to grow in stage 3, double-check here to ensure growth.

27. *Quality check the models.* Following the instructions in the original Protocol, perform quality check on the models generated including their mass/charge balance, dead-end metabolites/reactions and blocked reactions, etc.?Troubleshooting

28. *Validate strain-specific models.* Perform experiments or collect experimental data from the literature on the metabolic capabilities of the strains of interest. Data useful for validation include known secretion products, growth on different nutrient sources, auxotrophy and knock-out phenotypes (see original Protocol Steps 81 and 82 for details of model validation against experimental observations[6]). As with all GEMs, better experimental characterization of the strains of interest will improve the *in silico* results. Thus, increasing the accuracy of the biochemical composition of the biomass function for strains of interest is of value. !CAUTION: In order to maximize the accuracy of model prediction, ensure the simulation condition (constraint, strain, media) is consistent with the experimental condition.

5.9 Timing

The timing of the entire process is estimated under the assumption that the user has basic coding experience and is working with prokaryotes. The timing also depends on multiple factors: 1) Availability of the base model: the timing will be significantly reduced if the user starts with an available and high-quality base model. 2) Number of strains. While a good portion of the workflow can be automated (see Supplementary Tutorial), manual curation is still necessary for

each strain-specific model, resulting in longer time needed with increased number of strains. 3) Experience with coding/GEMs. If the user has worked with GEMs and is comfortable with coding (especially python), the timing will be greatly reduced with the help of the Supplementary Tutorial. 4) Computational resources. This factor will only come into play in the BLAST step if the user is working with large genomes and many strains. Otherwise a personal computer should be sufficient.

- Stage 1 (Steps 1-4) (base model reconstruction if model not available): 6 months - 1 year depending on the size of the genome, annotation quality and availability of the metabolic knowledge
- Stage 2 (Steps 5-13): days to weeks depending on the number of strains and availability of computational resources
- Stage 3 (Steps 14-23): days to weeks depending on the number of strains
- Stage 4 (Steps 24-28): days to weeks depending on the number of strains

5.10 Anticipated Results

This Protocol will result in multi-strain genome-scale metabolic models that not only can serve as a comprehensive knowledge base for the species of interest but will also allow computation of metabolic capabilities for different strains from just their genome sequences. Compared to single-strain-based GEMs, multi-strain GEMs can also be queried for strain-specific metabolic genes/reactions. Multi-strain GEMs will also allow various simulations including growth on different nutrient sources and gene knockouts to allow us to obtain a high-resolution understanding of the metabolic phenotypes displayed by different strains.

Acknowledgements

Chapter 5, in part, is a reprint of material published in: **Norsigian, Charles J.***, Xin Fang*, Yara Seif, Jonathan M. Monk, and Bernhard O. Palsson. "A workflow for generating multi-strain genome-scale metabolic models of prokaryotes." *Nature Protocols* (2019): 1-14. The dissertation author is one of the two primary authors.

5.11 References

1. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. en. *Curr. Opin. Genet. Dev.* **15,** 589–594 (Dec. 2005).

2. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. *Comparative genomics: the bacterial pan-genome* 2008.

3. Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V. & Ravel, J. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. en. *J. Bacteriol.* **190,** 6881–6893 (Oct. 2008).

4. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nat. Rev. Genet.* **15,** 107–120 (Feb. 2014).

5. 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).

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

7. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

8. Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L. & Palsson, B. O. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. en. *BMC Syst. Biol.* **12,** 66 (June 2018).

9. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

10. Norsigian, C. J., Kavvas, E., Seif, Y., Palsson, B. O. & Monk, J. M. iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of Acinetobacter baumannii AYE. en. *Front. Genet.* **9,** 121 (Apr. 2018).

11. Fouts, D. E., Matthias, M. A., Adhikarla, H., Adler, B., Amorim-Santos, L., Berg, D. E., Bulach, D., Buschiazzo, A., Chang, Y.-F., Galloway, R. L., Haake, D. A., Haft, D. H., Hartskeerl, R., Ko, A. I., Levett, P. N., Matsunaga, J., Mechaly, A. E., Monk, J. M., Nascimento, A. L. T., Nelson, K. E., Palsson, B., Peacock, S. J., Picardeau, M., Ricaldi, J. N., Thaipandungpanit, J., Wunder, E. A., Frank Yang, X., Zhang, J.-J. & Vinetz, J. M. *What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira* 2016.

12. 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. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

13. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. en. *Nucleic Acids Res.* **46,** 7542–7553 (Sept. 2018).

14. Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Murphy-Olson, D., Chan, S. Y., Kamimura, R. T., Kumari, S., Drake, M. M., Brettin, T. S., Glass, E. M., Chivian, D., Gunter, D., Weston, D. J., Allen, B. H., Baumohl, J., Best, A. A., Bowen, B., Brenner, S. E., Bun, C. C., Chandonia, J.-M., Chia, J.-M., Colasanti, R., Conrad, N., Davis, J. J., Davison, B. H., DeJongh, M., Devoid, S., Dietrich, E., Dubchak, I., Edirisinghe, J. N., Fang, G., Faria, J. P., Frybarger, P. M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J., Haun, H. L., He, F., Jain, R., Joachimiak, M. P., Keegan, K. P., Kondo, S., Kumar, V., Land, M. L., Meyer, F., Mills, M., Novichkov, P. S., Oh, T., Olsen, G. J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S. S., Price, G. A., Ramakrishnan, S., Ranjan, P., Ronald, P. C., Schatz, M. C., Seaver, S. M. D., Shukla, M., Sutormin, R. A., Syed, M. H., Thomason, J., Tintle, N. L., Wang, D., Xia, F., Yoo, H., Yoo, S. & Yu, D. KBase: The United States Department of Energy Systems Biology Knowledgebase. en. *Nat. Biotechnol.* **36,** 566–569 (July 2018).

15. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. en. *Nat. Biotechnol.* **32,** 447–452 (May 2014).

16. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

17. Fang, X., Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., Gianetto-Hill, C., Leung, N., Szubin, R., Sanders, J., Beck, P. L., Li, W., Sandborn, W. J., Gray-Owen, S. D., Knight, R., Allen-Vercoe, E., Palsson, B. O. & Smarr, L. Metagenomics-Based, Strain-Level Analysis of Escherichia coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient. en. *Front. Microbiol.* **9,** 2559 (Oct. 2018).

18. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic Acids Res.* **44,** D515–22 (Jan. 2016).

19. Ganter, M., Bernard, T., Moretti, S., Stelling, J. & Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. en. *Bioinformatics* **29,** 815–816 (Mar. 2013).

20. Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., *et al.* BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.* **34,** D689–D691 (2006).

21. Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Chauhan, S., Correia, K., *et al. Memote: A community driven effort towards a standardized genome-scale metabolic model test suite. bioRxiv. 2018; 350991*

22. Edwards, D. J. & Holt, K. E. *Beginner's guide to comparative bacterial genome analysis using next-generation sequence data, Microb. Inform. Exp. 3 (2013) 2*

23. Van Domselaar, G. H., Stothard, P., Shrivastava, S., Cruz, J. A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R. & Wishart, D. S. BASys: a web server for automated bacterial genome annotation. en. *Nucleic Acids Res.* **33,** W455–9 (July 2005).

24. Tanizawa, Y., Fujisawa, T. & Nakamura, Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. en. *Bioinformatics* **34,** 1037–1039 (Mar. 2018).

25. Seemann, T. Prokka: rapid prokaryotic genome annotation. en. *Bioinformatics* **30,** 2068–2069 (July 2014).

26. 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).

27. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E. & Ye, J. Database resources of the National Center for Biotechnology Information. en. *Nucleic Acids Res.* **40,** D13–25 (Jan. 2012).

28. 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).

29. 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).

30. Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., Magnusdóttir, S., Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Clancy, C. M., Modamio, J., Sauls, J. T., Noronha, A., Bordbar, A., Cousins, B., El Assal, D. C., Valcarcel, L. V., Apaolaza, I., Ghaderi, S., Ahookhosh, M., Ben Guebila, M., Kostromins, A., Sompairac, N., Le, H. M., Ma, D., Sun, Y., Wang, L., Yurkovich, J. T., Oliveira, M. A. P., Vuong, P. T., El Assal, L. P., Kuperstein, I., Zinovyev, A., Scott Hinton, H., Bryant, W. A., Aragón Artacho, F. J., Planes, F. J., Stalidzans, E., Maass, A., Vempala, S., Hucka, M., Saunders, M. A., Maranas, C. D., Lewis, N. E., Sauter, T., Palsson, B. Ø., Thiele, I. & Fleming, R. M. T. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. arXiv: `1710.04038 [q-bio.QM]` (Oct. 2017).

31. Palsson, B. Ø. *Systems Biology: Constraint-based Reconstruction and Analysis* en (Cambridge University Press, Jan. 2015).

32. Orth, J. D. & Palsson, B. Ø. Systematizing the generation of missing metabolic knowledge. en. *Biotechnol. Bioeng.* **107,** 403–412 (Oct. 2010).

33. Pan, S. & Reed, J. L. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. en. *Curr. Opin. Biotechnol.* **51,** 103–108 (June 2018).

34. Karp, P. D., Weaver, D. & Latendresse, M. How accurate is automated gap filling of metabolic models? en. *BMC Syst. Biol.* **12,** 73 (June 2018).

35. Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S. & Palsson, B. O. Systems approach to refining genome annotation. en. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 17480–17484 (Nov. 2006).

36. Beck, A. E., Hunt, K. A. & Carlson, R. P. Measuring Cellular Biomass Composition for Computational Biology Applications. en. *Processes* **6,** 38 (Apr. 2018).

37. Ekblom, R. & Wolf, J. B. W. *A field guide to whole-genome sequencing, assembly and annotation* 2014.

38. Angel, V. D. D., Del Angel, V. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M. & Lantz, H. *Ten steps to get started in Genome Assembly and Annotation* 2018.

39. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T. & Thiele, I. Generation

of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. en. *Nat. Biotechnol.* **35,** 81–89 (Jan. 2017).

40. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. en. *Nat. Rev. Microbiol.* **7,** 129–143 (Feb. 2009).

41. Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B. & Lercher, M. J. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. en. *PLoS Comput. Biol.* **13,** e1005494 (Apr. 2017).

42. Christopher Frey, H. & Patil, S. R. Identification and Review of Sensitivity Analysis Methods. *Risk Anal.* **22,** 553–578 (June 2002).

43. Schmelling, N. *Reciprocal Best Hit BLAST v1 (protocols.io.grnbv5e)*

44. Hulsen, T., Huynen, M. A., de Vlieg, J. & Groenen, P. M. A. Benchmarking ortholog identification methods using functional genomics data. en. *Genome Biol.* **7,** R31 (Apr. 2006).

45. Lachance, J.-C., Lloyd, C. J., Monk, J. M., Yang, L., Sastry, A. V., Seif, Y., Palsson, B. O., Rodrigue, S., Feist, A. M., King, Z. A. & Jacques, P.-É. *BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data* 2019.

Chapter 6

BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree

6.1 Abstract

The BiGG Models knowledge base (`http://bigg.ucsd.edu`) is a centralized repository for high-quality genome-scale metabolic models. For the past 12 years, the website has allowed users to browse and search metabolic models. Within this update, we detail new content and features in the repository, continuing the original effort to connect each model to genome annotations and external databases as well as standardization of reactions and metabolites. We describe the addition of 31 new models that expand the portion of the phylogenetic tree covered by BiGG Models. We also describe new functionality for hosting multi-strain models, which have proven to

be insightful in a variety of studies centered on comparisons of related strains. Finally, the models in the knowledge base have been benchmarked using Memote, a new community-developed validator for genome-scale models to demonstrate the improving quality and transparency of model content in BiGG Models.

6.2 Introduction

BiGG Models (http://bigg.ucsd.edu) was initially released in 2010 as a knowledge base of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions, and the first release was followed by a complete redesign in 2016 [1, 2]. Since its initial release, the BiGG Models publications have been cited over 450 times (via Web of Science) and the website maintains a user base of 2,000 monthly active users. BiGG Models is built around a workflow for standardizing models that is meant to verify and, in some cases improve, model quality. External studies have also indicated the high quality of models in BiGG. In one instance, the robustness of growth predictions for models in BiGG was demonstrated and used as a benchmark for a new collection of microbiome metabolic models [3]. Another study on "erroneous energy generating cycles"—a common issue in metabolic models—found that models in BiGG were less likely to have these undesirable cycles than models from other databases [4]. And a number of projects have used BiGG to automate reconstruction workflows and analyses [5–7].

With the BiGG Models 2020 update, we have included an additional 31 genome-scale metabolic models (GEMs) across four independent releases (versions 1.3–1.6), introduced the ability to download sets of multi-strain models that have been generated from a given base reconstruction page, and continuously improved features with suggestions and contributions from

the open source community. New content has increased the utility of the knowledge base for the community by expanding the number of organisms and metabolic processes represented. The BiGG Models architecture has been designed to enable these advances and continually improve the knowledge base.

6.3 Knowledge Base Content

BiGG Models continues to contain high-quality, manually-curated GEMs collected from various publications. Quality control in BiGG Models begins with our requirement that all models undergo rigorous peer review before entry. We begin our import workflow with the exact model that was reported in a peer-reviewed publication, and the workflow is designed to improve the quality of annotations and standardization in the model, without making any changes to the reaction content, parameterization, or relationships (e.g. gene-reaction rules).

To load a model into BiGG, first each model is aligned the shared namespace of reactions and metabolites across all models. When identifiers can be improved automatically (e.g. by finding a universal reaction based on the reactants), the workflow does this automatically; in other cases, non-matching identifiers are left as-is to ensure that model content does not change. Next, genome annotations are loaded into the database for each model, providing explicit links between metabolic reactions and genes. When adding content to the BiGG Models database, manual efforts are made to ensure that each metabolite identifier follows the specified naming convention, each reaction contains a unique identifier, and gene reaction rules are properly represented in valid Boolean logic. When obvious errors are identified (typos, duplicate metabolites), these are corrected manually, with feedback from the model authors. The coalescence of genome annotation information, with external database links, and reaction, metabolite, gene information

from peer-reviewed models drive the quality of the knowledge base.

To ensure that model content (the reaction connectivity, gene reaction rules, and parameters that affect model predictions) has not changed from the peer-reviewed version presented in the original publication, an internal testing suite runs 18 tests for each model, for a total of more than 1900 tests. For example, tests ensure that reaction, metabolite, and gene counts have have not changed, that all reactions that were mass balanced in the published model are still balanced, and that genes have mapped to genome annotations correctly. An additional 36 tests are included to spot-check bugs and edge cases that have appeared during previous builds of BiGG Models. The full test suite is available in the source code (`https://github.com/SBRG/bigg_models/blob/master/bigg_models/tests`).

In the 2016 release of BiGG Models, there were 77 GEMs; with this update, we detail 31 additional models, covering release versions 1.3–1.6 (`http://bigg.ucsd.edu/updates`)[8–13]. Genome annotations for each model (where possible) are downloaded from the National Center for Biotechnology (NCBI) reference sequence database [14] and linked to the corresponding GEM. Notable additions are the Recon3D, iCHOv1, and iML1515 [15–17] for the human metabolic network, Chinese hamster ovary cell, and Escherichia coli K-12 MG1655 respectively. BiGG Models continues to host gold-standard models within a shared knowledge base of biological reactions and metabolites. We also demonstrate that the new GEMs valuably expand the portion of the reactome encapsulated by the knowledge base. The number of reactions represented in the database more than doubled from 11,459 in the 2016 version to 28,302. Likewise, the number of metabolites has more than doubled from 4,040 to 9,088. In addition to expanding the number of metabolic processes within the database, we sought to evaluate the diversity of reaction presence among GEMs within the database. Reaction presence or absence of the shared namespace was

identified for every representative GEM, and this matrix was subject to multiple correspondence analysis (Figure 6.1). Notably, this analysis shows that new models within the update exist at the edge of each cluster demonstrating that the new content is increasing the level of dissimilarity amongst GEM reaction content. This separation among models conveys that the metabolic space within BiGG Models is moving past representations of shared common pathways and incorporating an increasing amount of organism-specific biochemical capabilities.

This update also includes multi-strain models, a recent development within the metabolic modeling community. We define multi-strain models as those generated via the ability to extend the content contained within a gold-standard reconstruction to related strains of interest. This technique has proven insightful in a number of studies for comparative analysis of strains [18–24]. Thus, we have included a means for the hosting of the draft strain-specific models generated within these studies on BiGG Models. Each strain-specific model is available to download within a zip folder from the page of the base reconstruction used to generate the strain-specific models. The GEMs of iCN718, iYL1228, and STM_v1_0 [18, 25, 26] each contains datasets of multi-strain models linked from their reconstruction pages within BiGG Models. Identifiers in multi-strain reconstructions are inherently BiGG Models compliant as they have been generated through the use of a hosted model. These multi-strain models have demonstrated value in comparative simulation to identify key differences amongst the strains of a species and they all represent starting points towards manually curated reconstructions for each strain should the proper steps be undertaken [27].

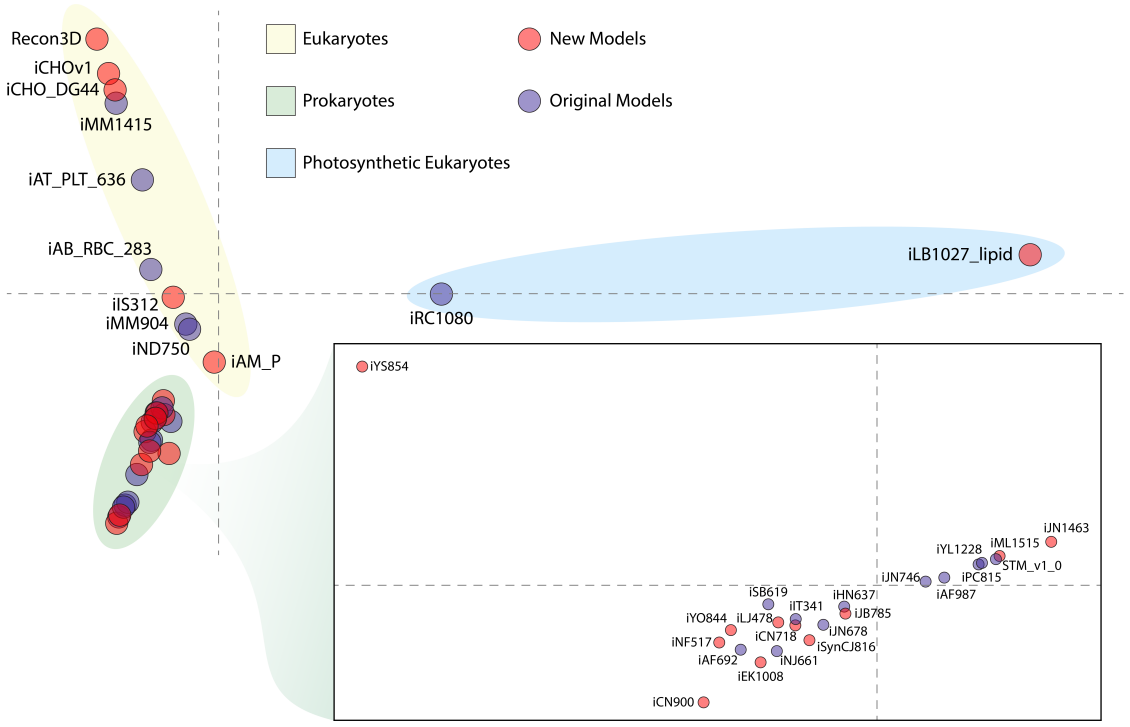

Figure 6.1: Multiple correspondence analysis of the reaction presence or absence within each model clusters models according to eukaryotic (yellow ellipse), prokaryotic (green ellipse and inset), and photosynthetic eukaryotes (blue ellipse) within metabolic reaction space. Dimension 1 (x-axis) explained 14.5% of the variance; dimension 2 (y-axis) explained 14.2%. Further, a number of the models newly introduced within this update (red circles) are found edges of the MCA plot, indicating that within these two dimensions, they contribute to additional diversity in reaction content compared to the previous release. For this analysis iML1515 was used as a representative *E. coli* model and iIS312 as representative for *Trypanosoma cruzi*.

6.4 Validation of Models with Memote

BiGG Models now links to the model validation tool, Memote, which evaluates and scores GEMs with a set of community-maintained tests [28]. Consistent with the efforts in BiGG Models to maximize the value of metabolic models, evaluation with Memote provides a means to quantify model quality. Quality, in this case, indicates that GEMs adhere to established standards such as consistent identification of model components and biologically-feasible results under varied growth conditions. This standardized approach to model validation ensures the quality of BiGG Models content and provides a benchmark for continued improvement.

Both the original 77 GEMs included in the 2016 release of BiGG and the 31 GEMs included in this update were evaluated with Memote (Figure 6.2). Largely due to improved gene, metabolite, and reaction annotations, the average Memote score of JSON formatted models increased from 40% to 58%, while that of the SBML [29–31] formatted models advanced from 66% to 73%. While these scores represent significant improvements, ongoing database annotation efforts will be necessary to maximize Memote scores for models in BiGG. Memote does not currently support testing of MATLAB formatted models; however, BiGG generates MATLAB-formatted models using the same data sources as the JSON formatted files, so equivalent model content is present. These results highlight the value of BiGG Models as a knowledge base of GEMs, and scoring its content with Memote reinforces its effort to provide access to GEMs with thorough and consistent standards.

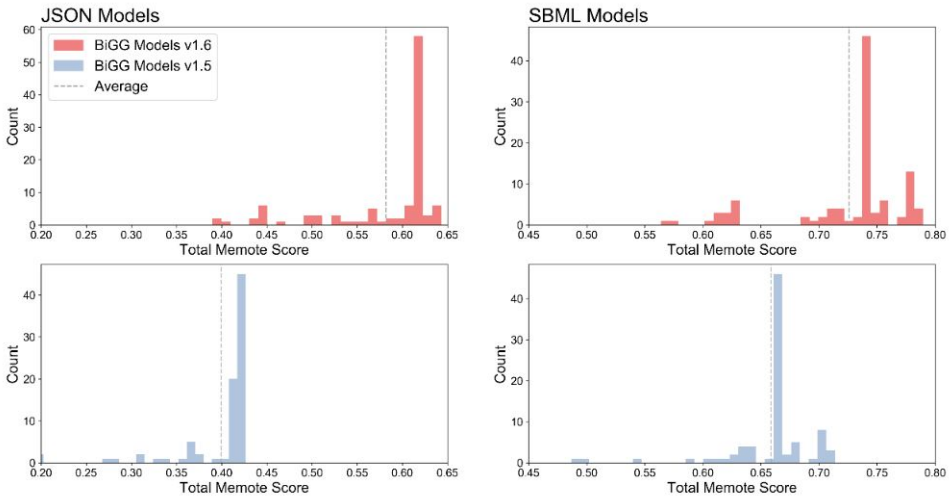

Figure 6.2: The latest update has resulted in improved Memote annotation scores for both JSON and SBML model formats.

6.5 Additional Features and Improvements

Regular improvements are made to BiGG Models that have made the knowledgebase faster, easier to use, and better for analysis. Filters are now provided during search to filter out multi-strain reconstructions in the search results (see the toggle titled "Exclude multistrain models from search"). Gene and protein sequences are now included directly in the database and available by API. A new advanced search feature allows users to identify all gene and protein sequences for any universal BiGG reaction (see "Find sequences for BiGG Models reaction" on the advanced search page).

A new "universal" model was added for download on the Data Access page; this model provides all reactions and metabolites from BiGG in a single COBRA-compatible JSON file, so users can rapidly add BiGG content to their own computational workflows using COBRA tools. Namespace downloads on the Data Access page have also been extended to include old and deprecated identifiers. External database links are regularly updated with the latest information from MetaNetX [32]. Many manual improvements have been made to annotations, including better gene mapping for yeast models. And SBML downloads have improved through regular updates to the ModelPolisher project (`https://github.com/draeger-lab/ModelPolisher`).

Since the 2016 release of BiGG Models, the website has been deployed on a new server to dramatically improve speed when searching and browsing. Finally, bugs and suggestions are collected on GitHub (`https://github.com/SBRG/bigg_models`), and this has led to continuous and transparent improvements to the site by the BiGG Models team.

6.6 Conclusion

BiGG Models continues to be a widely used and well-maintained platform for integrating, sharing and standardizing GEMs. The updated knowledge base integrates the metabolic knowledge for 108 GEMs, as well as including the content for 515 draft strain-specific models across three organisms, all available within the knowledge base. BiGG Models is free for academic use and continues to extend the content within the knowledge base. Further, all source code continues to be available on GitHub to enable submission of potential bugs. The development of BiGG Models continues to evolve with the needs of the research community, introducing multi-strain models and validation through Memote testing. Future BiGG Models releases will continue to be shaped by the feedback from users.

6.7 Data Availability and Requirements

BiGG Models is freely available online for academic and non-profit use at `http://bigg.ucsd.edu`, under the BiGG License described at `http://bigg.ucsd.edu/license`. While the content of BiGG is restricted to academic and non-profit use to protect intellectual property claims, the source code is open source and available to all users under the MIT license at `https://github.com/SBRG/bigg_models`. Installation of an independent system requires Python 3.5 and PostgreSQL 9.4 or later.

We encourage community members to submit their model content to BiGG Models, and the website includes a section that describes the minimum requirements for inclusion in BiGG and the process for submitting a new model: `http://bigg.ucsd.edu/about` These requirements reflect the quality standards set by BiGG Models: identifier standardization for reactions and

metabolites, links to genome annotations and peer-reviewed publication as the primary means of verifying model quality.

Acknowledgements

Chapter 6, in part, is a reprint of material published in: **Norsigian, Charles J.**, Neha Pusarla, John Luke McConn, James T. Yurkovich, Andreas Dräger, Bernhard O. Palsson, and Zachary King. "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree." *Nucleic acids research* 48, no. D1 (2020): D402-D406. The dissertation author was the primary author.

6.8 References

1. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. en. *BMC Bioinformatics* **11,** 213 (Apr. 2010).

2. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic Acids Res.* **44,** D515–22 (Jan. 2016).

3. Babaei, P., Shoaie, S., Ji, B. & Nielsen, J. Challenges in modeling the human gut microbiome. en. *Nat. Biotechnol.* **36,** 682–686 (Aug. 2018).

4. Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B. & Lercher, M. J. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. en. *PLoS Comput. Biol.* **13,** e1005494 (Apr. 2017).

5. Chan, S. H. J., Cai, J., Wang, L., Simons-Senftle, M. N. & Maranas, C. D. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. en. *Bioinformatics* **33,** 3603–3609 (Nov. 2017).

6. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. en. *Nucleic Acids Res.* **46,** 7542–7553 (Sept. 2018).

7. Xavier, J. C., Patil, K. R. & Rocha, I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. en. *Metab. Eng.* **39,** 200–208 (Jan. 2017).

8. Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., Golden, S. S. & Palsson, B. O. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E8344–E8353 (Dec. 2016).

9. Levering, J., Broddrick, J., Dupont, C. L., Peers, G., Beeri, K., Mayers, J., Gallina, A. A., Allen, A. E., Palsson, B. O. & Zengler, K. Genome-Scale Model Reveals Metabolic Basis of Biomass Partitioning in a Model Diatom. en. *PLoS One* **11,** e0155038 (May 2016).

10. Calmels, C., McCann, A., Malphettes, L. & Andersen, M. R. Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. en. *Metab. Eng.* **51,** 9–19 (Jan. 2019).

11. Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J. & Feist, A. M. Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. en. *Cell Syst* **3,** 238–251.e12 (Sept. 2016).

12. 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).

13. Abdel-Haleem, A. M., Hefzi, H., Mineta, K., Gao, X., Gojobori, T., Palsson, B. O., Lewis, N. E. & Jamshidi, N. Functional interrogation of Plasmodium genus metabolism identifies species- and stage-specific differences in nutrient essentiality and drug targeting. en. *PLoS Comput. Biol.* **14,** e1005895 (Jan. 2018).

14. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E. & Ye, J. Database resources of the National Center for Biotechnology Information. en. *Nucleic Acids Res.* **40,** D13–25 (Jan. 2012).

15. Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., Prlić, A., Sastry, A., Danielsdottir, A. D., Heinken, A., Noronha, A., Rose, P. W., Burley, S. K., Fleming, R. M. T., Nielsen, J., Thiele, I. & Palsson, B. O. Recon3D enables a three-dimensional view of gene variation in human metabolism. en. *Nat. Biotechnol.* **36,** 272–281 (Mar. 2018).

16. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

17. Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., Paglia, G., Loira, N., Spahn, P. N., Pedersen, L. E., Gutierrez, J. M., King, Z. A., Lund, A. M., Nagarajan, H., Thomas, A., Abdel-Haleem, A. M., Zanghellini, J., Kildegaard, H. F., Voldborg, B. G., Gerdtzen, Z. P., Betenbaugh, M. J., Palsson, B. O., Andersen, M. R., Nielsen, L. K., Borth, N., Lee, D.-Y. & Lewis, N. E. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. en. *Cell Syst* **3,** 434–443.e8 (Nov. 2016).

18. Norsigian, C. J., Kavvas, E., Seif, Y., Palsson, B. O. & Monk, J. M. iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of Acinetobacter baumannii AYE. en. *Front. Genet.* **9,** 121 (Apr. 2018).

19. Norsigian, C. J., Attia, H., Szubin, R., Yassin, A. S., Palsson, B. Ø., Aziz, R. K. & Monk, J. M. Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase-Producing Multidrug-Resistant Klebsiella pneumoniae Clinical Isolates. en. *Front. Cell. Infect. Microbiol.* **9,** 161 (May 2019).

20. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

21. Fouts, D. E., Matthias, M. A., Adhikarla, H., Adler, B., Amorim-Santos, L., Berg, D. E., Bulach, D., Buschiazzo, A., Chang, Y.-F., Galloway, R. L., Haake, D. A., Haft, D. H., Hartskeerl, R., Ko, A. I., Levett, P. N., Matsunaga, J., Mechaly, A. E., Monk, J. M., Nascimento, A. L. T., Nelson, K. E., Palsson, B., Peacock, S. J., Picardeau, M., Ricaldi, J. N., Thaipandungpanit, J., Wunder Jr, E. A., Yang, X. F., Zhang, J.-J. & Vinetz, J. M. What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira. en. *PLoS Negl. Trop. Dis.* **10,** e0004403 (Feb. 2016).

22. Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L. & Palsson, B. O. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. en. *BMC Syst. Biol.* **12,** 66 (June 2018).

23. 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. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

24. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia

coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

25. Thiele, I., Hyduke, D. R., Steeb, B., Fankam, G., Allen, D. K., Bazzani, S., Charusanti, P., Chen, F.-C., Fleming, R. M. T., Hsiung, C. A., De Keersmaecker, S. C. J., Liao, Y.-C., Marchal, K., Mo, M. L., Özdemir, E., Raghunathan, A., Reed, J. L., Shin, S.-I., Sigurbjörnsdóttir, S., Steinmann, J., Sudarsan, S., Swainston, N., Thijs, I. M., Zengler, K., Palsson, B. O., Adkins, J. N. & Bumann, D. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. en. *BMC Syst. Biol.* **5,** 8 (Jan. 2011).

26. Liao, Y.-C., Huang, T.-W., Chen, F.-C., Charusanti, P., Hong, J. S. J., Chang, H.-Y., Tsai, S.-F., Palsson, B. O. & Hsiung, C. A. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. en. *J. Bacteriol.* **193,** 1710–1717 (Apr. 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. Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Chauhan, S., Correia, K., *et al. Memote: A community driven effort towards a standardized genome-scale metabolic model test suite. bioRxiv. 2018; 350991*

29. Hucka, M., Bergmann, F. T., Hoops, S., Keating, S. M., Sahle, S., Schaff, J. C., Smith, L. P. & Wilkinson, D. J. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core. en. *J. Integr. Bioinform.* **12,** 266 (Sept. 2015).

30. Olivier, B. G. & Bergmann, F. T. SBML Level 3 Package: Flux Balance Constraints version 2. en. *J. Integr. Bioinform.* **15** (Mar. 2018).

31. Hucka, M. & Smith, L. P. SBML Level 3 package: Groups, Version 1 Release 1. en. *J. Integr. Bioinform.* **13,** 290 (Dec. 2016).

32. Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A. & Pagni, M. MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44,** D523–D526 (2015).

Chapter 7

Systems biology approach to functionally assess the *Clostridioides difficile* pan-genome reveals genetic diversity with discriminatory power

7.1 Abstract

Combatting *Clostridioides difficile* infections, a dominant cause of hospital associated infections with incidence and resulting deaths increasing worldwide, is complicated by the frequent emergence of new virulent strains. Here we employ whole genome sequencing, high throughput phenotypic screenings and genome-scale models of metabolism to evaluate factors underlying *C. difficile* strain emergence by analyzing the genetic and phenotypic diversity of 451 strains. Con-

structing the *C. difficile* pan-genome based on this set revealed 9,924 distinct gene clusters of which 2,899 (29%) are defined as core, 2,968 (30%) are defined as unique and the remaining 4,057 (41%) are defined as accessory. We develop a novel strain typing method, Sequence Typing by Accessory Genome (STAG), that identifies 176 genetically distinct groups of strains and allows for explicit interrogation of accessory gene content. Thirty-five strains representative of the overall set were experimentally profiled on 95 different nutrient sources revealing 26 distinct growth profiles and unique nutrient preferences. Strain-specific genome scale models of metabolism were constructed for each of the strains to mechanistically link the observed phenotypes to strain-specific genetic differences exhibiting an ability to correctly predict growth in 76% of cases. The typing and model predictions are used to identify and contextualize relevant genetic features and phenotypes that may contribute to the emergence of new problematic strains.

7.2 Introduction

Clostridioides difficile remains the most common healthcare-associated infection with an ever-evolving and complex epidemiology. *C. difficile* is recognized as an urgent threat by the Centers for Disease Control and Prevention (CDC) and has been conservatively estimated at over 220,000 cases in hospitalized patients and nearly 13,000 deaths within the United States annually [1]. The disruption of natural colonic microbiota following antibiotic use is the leading risk factor for *C. difficile* infection (CDI) and recurrent infections occur in 35% of patients [2–4]. Two toxins, TcdA and TcdB, are the primary virulence factors for symptomatic infection [5]. However, virulence is also attributed by other factors including the cytolethal distending toxin (CDT), sporulation, flagella, and adhesins [6–12]. Overall, the plasticity of the *C. difficile* genome has contributed to divergent lineages distinguished by evolutionarily advantageous genetic traits

that result in increased antimicrobial resistance, virulence, and metabolic capabilities for survival within the gut [13, 14]. The bevy of accessory gene content present across strains in this species has complicated attempts to contextualize strain relationships amongst this complex population.

Molecular typing techniques that evaluate strain relatedness have been used to evaluate *C. difficile* epidemiology and track transmission of virulent lineages. The *C. difficile* genome has sufficient intraspecies diversity within the intergenic spacer regions of rRNA genes for the successful use and adoption of PCR ribotyping, the primary molecular typing method for C. difficile [15–18]. As a result, the most prevalent and hypervirulent *C. difficile* strains globally have been dubbed ribotype 027 (RT027) and ribotype 078 (RT078) [12, 19, 20]. Additionally, multilocus sequence typing (MLST) is widely used in population studies as a means of distinguishing strains through the allelic profile of designated housekeeping genes [21–23]. In addition to these two techniques there are several other typing methods including multilocus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, restriction endonuclease analysis, toxinotyping, and surface-layer protein A-encoding gene typing. Each of these methods has unique levels of discriminatory power as well as unique limitations [24]. While these typing schemes have proven useful in understanding CDI epidemiology, the most widely adopted schemes (PCR ribotyping and MSLT) lack the resolution to distinguish more closely related strains. Further to obtain mechanistic insight into outbreaks, whole genome sequencing (WGS) methods need to be employed.

Advancements in sequencing technologies have resulted in an explosion in the availability of quality whole-genome sequencing data [25] promising new and comprehensive approaches to strain typing [26–28]. In this age of high-throughput sequencing, comparative genomics analysis has been largely stratified into two approaches: single nucleotide variants (SNVs) and gene-by-

gene comparisons. In the later case for *C. difficile*, core-genome MLST (cgMLST) and whole-genome MLST (wgMLST) extensions of classical MLST have been developed [29, 30]. While these techniques have increased the resolution of typing approaches, key connections between the genomic diversity driving strain types and resulting diversity of phenotypes have remained elusive. A deeper understanding of the functional diversity across this species is needed and must be rooted to the enormous genetic diversity observed.

In recent years, systems biology tools have been challenged with extracting knowledge from the enormous amount of omics data available. In particular the substantial variability in genomic content and function across strains of a species can be analyzed efficiently through a combination of comparative genomics and various modeling frameworks [31–33]. Strain-specific genetic variation can be usefully organized through a pan-genomic perspective that delineates and organizes a species' gene portfolio [34, 35]. Additionally, genome-scale models (GEMs) of metabolism have served as tools to mechanistically link genotype to phenotype particularly in terms of growth capabilities. Computation of catabolic capabilities based on genome sequences has provided additional insight into metabolic variability and association to lifestyle niche [36, 37]. To increase understanding of the diversity exhibited by *C. difficile*, we have executed a holistic systems biology analysis encompassing both a functional genomics assessment of the pan-genome and an in depth analysis of experimental growth phenotypes aided by construction and use of GEMs. Moreover, we developed a novel strain typing method based on the accessory gene content, Sequence Typing by Accessory Genome (STAG), that allows for explicit investigation into the gene clusters driving the separation of strain groups. This new method expands the toolkit for analysis of WGS strain typing across a broad array of disciplines.

7.3 Results

7.3.1 High-throughput phenotypic screening of *C. difficile* clinical isolates reveals unique dynamic growth profiles

To evaluate the metabolic capabilities of *C. difficile*, we profiled 35 clinical strains isolated from hospitalized adult patients [38] using Biolog Phenotype Microarrays and evaluated their ability to catabolize 95 unique carbon sources (Methods). Analysis of the time-course data demonstrated various growth modalities (Figure 7.1A). Gaussian process regression models were employed to robustly explore these dynamics (Methods). Inferring growth curves and their time derivatives from our data enables the calculation of traditional growth model parameters such as carrying capacity (K), maximum growth rate, doubling time, and area under the curve (AUC) through a non-parametric approach [39]. Gaussian process (GP) regression is advantageous because it has been shown to outperform parametric approaches when considering non-traditional growth-curve shapes such as diauxic shifts and long lag phases [40, 41]. Examination of the primary growth model parameters demonstrated that the carrying capacity and area under the curve are the best indicators of binary microbial growth. The K value represents the maximum population size and the AUC value is a measurement of net growth over time irrespective of curve shape, therefore these metrics are particularly suited for high throughput screens of divergent strains. Following sensitivity analysis (Methods), we normalized values to the negative control and determined that AUC value greater than 1.25 and a K value greater than .3 define a high confidence growth call (Figure 7.1B). This combined threshold increased confidence of growth analysis by minimizing impact of data noise and growth dynamics as compared to simple fold change or maximum optical density thresholds [42].

Overall, unanimous growth determinations (either growth supporting or non-growth supporting) could be made for 67 compounds, 4 (glucose, fructose, mannitol, n-acetyl-d-glucosamine) of which were universally growth-supporting across the 35 strains while the remaining 63 were unanimous non-growth supporting. The remaining 28 carbon sources assayed support growth in a range of one to 34 strains. Therefore, these 28 carbon sources could be used to construct an overall metabolic profile encompassing the growth capabilities on each of these substrates (Supplementary Figure E.1). For example, CDS031 was the only strain found to grow on galactose, while growth on sucrose was limited to strains CDS071 and CDS031. Niche growth capabilities are identified by examining the outliers in parameter values from the overall set (Figure 7.1C). In particular, the degree of growth support can be investigated through the calculated AUC and carrying capacity. Ranking calculated AUC and carrying capacity reveals which substrates are the strongest strain-specific growth supporters. Outside of the four universal growth supporting nutrients, the next top five substrates vary across the strains and include: mannose, sorbitol, trehalose, sucrose, maltose, glycerol, n-acetyl-d-mannosamine, serine and threonine (Supplementary Figure E.2). This data indicates that while serine supports growth of multiple strains, only CDS078 grows robustly on serine as one of its best substrates.

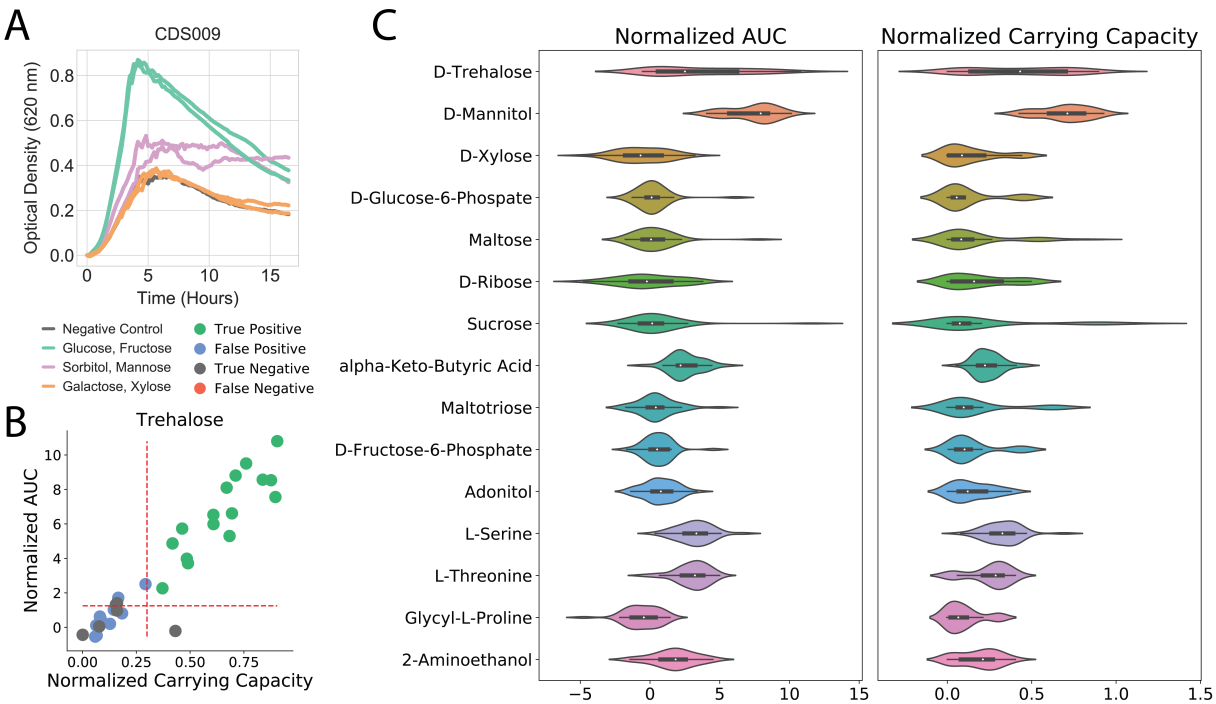

Figure 7.1: Growth dynamics of *C. difficile* isolates and parameters calculated through gaussian process regression. A) Growth curves for one isolate, CDS009, on 6 of 95 carbon substrates demonstrating variable growth dynamics and shape of growth curves. B) Each of the 35 isolates growth on trehalose plotted in AUC and K with thresholds of 1.25 and .3 shown as red dashed lines, strains are colored by corresponding genome scale model prediction of growth with experimental data. C) Of the 28 discriminatory carbon sources, the top 15 in terms of coefficient of variation of AUC and K between strains are pictured.

7.3.2 GEM-predicted capabilities capture discriminatory metabolic profiles

Motivated by the diverse catabolic capabilities identified through our metabolic profiling and subsequent GP regression modeling, we sought to identify the genetic bases for these different capabilities. Genome-scale models (GEMs), in particular multi-strain modeling, provide a powerful tool to contextualize genetic differences and generate metabolic predictions [36, 37, 43–45]. Therefore, we generated strain-specific GEMs for each one of our 35 isolates based on iCN900, a gold-standard reconstruction of *C. difficile* strain 630 [46, 47]. To facilitate generating GEMs of our 35 strains, we completed whole genome sequencing of each isolate and then executed a standard protocol to build draft strain-specific models based on the reference reconstruction iCN900 [48]. Our preliminary comparative genomics analyses using the reference 630 sequence (AM180355.1) are summarized through principal component analysis of shared genes across the entire genomes of our 35 strains (Figure 7.2A). This analysis demonstrates that the clinical isolates exhibit variations in conserved genes relative to the reference sequence and that this variation is not consistent across ribotypes.

We evaluated the conserved subsystems of metabolism across the models and found that transport functions, metabolism of particular amino acids, fatty acid metabolism, and starch and sucrose metabolism were most divergent against the reference amongst the strains (Figure 7.2B). Specifically the reactions of phosphotransferase system and ABC system transporters (86%), starch and sucrose metabolism (57%), fatty acid biosynthesis (21%), and lysine and arginine pathways (20.8%) have a high proportion of reactions whose encoding genes contain at least one non conserved gene (Supplementary Figure E.3). A major power of GEMs is their ability to predict phenotypes based on the structure of the metabolic network using flux balance analysis (FBA) [49, 50]. Thus, we used our strain-specific GEMs to generate model predictions for growth

on all 95 carbon sources contained within the phenotypic microarray growth data. In silico growth predictions were generated using previously defined minimal media conditions and alternating the carbon source (Methods). Each strain-specific model was subsequently individually gap-filled and specific false-negative model predictions offered opportunities for further curation (Methods). This led to the addition of reactions to specific strains that enabled in silico biomass production for growth on the following sole carbon sources: pyruvate, n-acetyl-D-mannosamine, D-fructose-6-phosphate, D-glucose-6-phosphate, D-serine, and maltotriose bringing these compounds into agreement with experimental profiles.

Critically, we compared the resulting confusion matrix between our experimental dataset and GEM model predictions (Figure 7.2C). Using the GP parameters as opposed to fold change to determine experimental binary growth/no growth calls elevates the cohort of curated strain-specific GEMs in overall accuracy by 10% and the Matthews Correlation Coefficient (MCC) of the predictions by 0.31, resulting in an overall accuracy of 76% and 0.41 MCC. This improvement demonstrates that by enacting a more stringent analysis of high-throughput experimental screening data, the accuracy of growth versus no growth calls can be improved. Importantly, these additional analyses minimized the impact of incidental spikes and fluctuations in optical density, which leads to faulty growth calls. Stringent growth calling of experimental data can also lead to a high degree of false-positive GEM predictions (765 for our dataset, Figure 7.2C), which usually occur because FBA simulation will find any theoretical solution possible dependent on network content and does not consider transcriptional regulation or enzyme efficiency [51]. This predictive failure mode in our set of models suggests that in addition to metabolic network diversity other biological processes play a role in the diverse capabilities of these strains. Thus we expanded our analysis from curated reactomes to a full pan-genome level analysis.

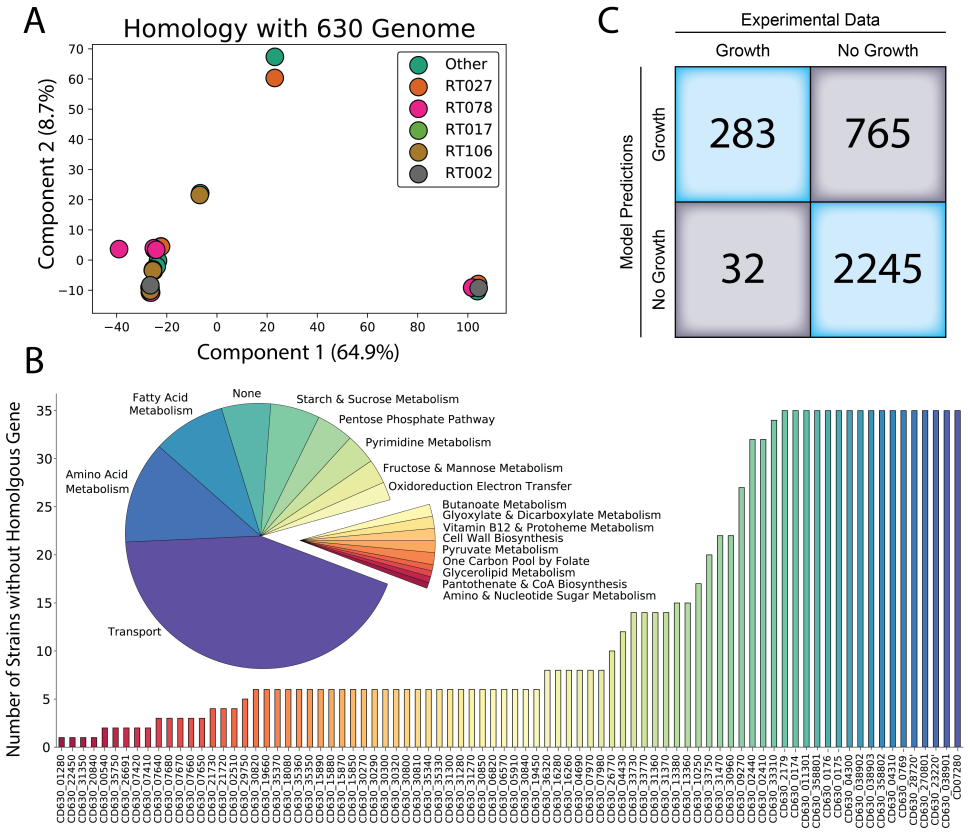

Figure 7.2: Whole genome similarity to reference strain 630, deviation in portion of gene portfolio contained within iCN900, and overall accuracy of 35 strain-specific models A) Principal component analysis of the matrix of whole genome homology of each isolate against *C. difficile* 630. Epidemic ribotypes are highlighted and represented in each cluster suggesting that their relationship to the reference strain is diverse across these lineages. B) Initial gene content removed from the set of 35 models based on lack of homologous genes from iCN900 and corresponding reaction metabolic subsystems. C) Final agreement of curated strain-specific isolate models and experimental profiling data resulting in 76% accurate set of 35 models.

7.3.3 Characterization of the *C. difficile* pan-genome demonstrates differences in conservation based on functional classification

To comprehensively analyze the diversity of strain-specific gene portfolios on a species level, we collected 416 high-quality, publicly available genomes. Along with our clinical isolate dataset this expanded our overall scope to 451 strains, which were all re-annotated to avoid potential biases from differential gene calling (Methods). We generated a phylogenomic tree for this dataset and examined how our clinical isolate genomes relate to the public dataset (Figure 7.3A). Our isolates cover 14 out of the 33 major tree branches and thus span 42% of the *C. difficile* phylogeny analyzed here.

To evaluate conserved and unique genes across the strains we constructed a pan-genome using the 451 genome sequences described above (Methods). The pan-genome is built through efficient all-by-all sequence homology comparisons (via CD-HIT) that establish gene clusters ranging from unique to ubiquitous genes. Our analysis identified a total of 9,924 gene clusters in the *C. difficile* pan-genome, where 2,899 are shared by 99% or more (446/451) of the strains and comprise the core-genome (Figure 7.3B). Likewise we identified 2,968 gene clusters present in only 1% or less (4/451) of the strains defining the unique-genome. The remaining 4,057 gene clusters represent the accessory-genome that is variably present within the population, but not present at either the core or unique extremes and therefore provide a genetic bank rich in discriminatory power.

The gene clusters were functionally annotated using EggNOG [52] and the results parsed into the broad category COGs: Metabolism, Cellular Processes and Signaling, and Information Storage and Processing (Figure 7.3C). Any COG assignment falling under "Poorly Characterized" were lumped into the genes with no annotation information to form the "Uncharacterized"

group (Figure 7.3D). Splitting the pan-genome into its functional constituents showed that genes with a Metabolic classification compose less accessory content and the genes encoding metabolic functions create the most closed pan-genome curve. This is in agreement with the high degree of false positive predictions made by our 35 strain-specific models as GEMs are predictors of what is feasible based on presence of encoding genes, but lack regulatory context for expression of those genes. Further, 68.3% of the overall pan-genome is classified as Uncharacterized and these gene clusters have the greatest accessory to core ratio and most open pan-genome curve, demonstrating the significant knowledge gaps still present for the species. To shed light on uncharacterized genes that may impact the measured metabolic phenotypes we calculated the biserial correlation between measured phenotypes and presence/absence of gene clusters. In total, 374 unique gene clusters were found to be positively correlated with one or more phenotypes at a p-value$<$0.001.

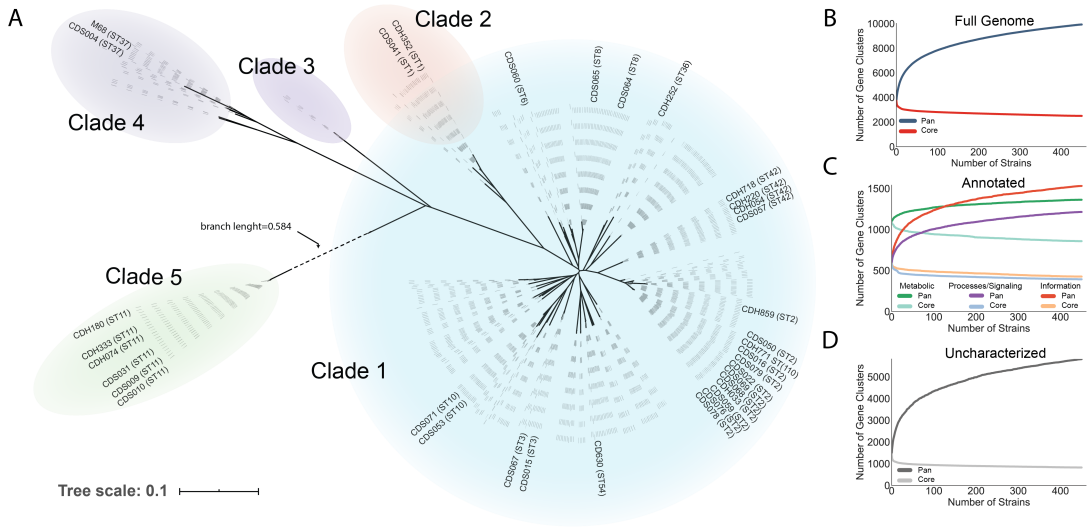

Figure 7.3: Phylogenomics and Pan and Core Genome curves for the 451 strain set. A) Phylogenomic tree constructed using 451 strains and clinical isolates labeled therein. Each dashed line represents one strain. B) Considering the totality of gene clusters, the core-genome is defined by the universally present 2,899 gene clusters and the remaining 7,025 gene clusters (accessory and unique) make up the rest of the pan-genome. C) For gene clusters where functional annotation can be assigned through COG categories, we analyzed the accessory/core breakdown of each major functionally defined group and the behavior of the pan-genome curves. D) The uncharacterized or annotated as "function unknown" clusters make up 68.37% of all gene clusters and these clusters exhibit the most open behavior in the pan-genome curve. This is indicative of the vast amount of *C. difficile* genes whose function remain unknown and present numerous candidates for discovery.

7.3.4 Functional assessment of accessory genome provides discriminatory power

We first evaluated the concordance between a SNP based phylogenetic tree and one created from a hierarchical clustering of the accessory genome represented in a binary format (Supplementary Figure E.4). We found that the trees had a correlation of 0.55 and entanglement of 0.12 indicating that accessory genome content is not completely concordant with SNP-based phylogeny. To evaluate the effect of this phenomena on MLST-defined sequence types we measured the association between accessory genome clusters and defined ST types using Cramer's V statistic. In total 9% of accessory gene families were highly associated with more than 1 ST (361 found in at least 2 ST types with V>0.4).

Based on this result we sought to develop an alternative strain typing scheme based on the accessory genome. The *C. difficile* community commonly uses approaches such as SNP trees, PCR-ribotyping and MLST types to distinguish strains. MLST and ribotyping have been shown to be similar in discriminatory capabilities, but do not have a direct one-to-one mapping classification of strains [21]. A pan-genome based strain typing scheme should resolve groups of strains within a species as well as provide the ability to interrogate the biological relevance of genetic drivers separating different groups. As strain-specific differences have been shown to be critical factors for differentiating phenotypes such as nutrient niches [43, 53], virulence [54–57], and antimicrobial resistance [58, 59], the ability to distinguish isolates from each other in a way that immediately assigns functional relevance will enhance global epidemiology. To this end, we introduce sequence typing by accessory genome (STAG), an algorithm that capitalizes on the untapped opportunity to classify strain groupings based on the diversity of the accessory gene portfolio.

The STAG algorithm utilizes accessory gene clusters to represent each genome as a binary profile that is defined by gene presence/absence within each accessory gene cluster (Supplementary Figure E.5). STAG then uses the Jaccard similarity index, defined as the size of intersection between two binary sets divided by the size of their union, to evaluate how similar each strain vector is to another [60]. Following the calculation of Jaccard similarity, STAG establishes a symmetric matrix composed of pairwise strain similarity, which is used to sort strains into groupings (Methods). Next, STAG incorporates the simple metric of compression factor to prioritize strain groupings, which we define as the number of strains divided by the number of groups. Briefly, STAG sorts strains into pan-genome types (PGTs) by iteratively passing over the similarity matrix checking for exclusive groupings based on a given threshold of similarity. At each pass the matrix is sorted according to a range of thresholds and the threshold that maximizes the compression factor of exclusive groups is selected for that pass. STAG removes the strains of exclusive groups as PGTs and the threshold identified is set as the new threshold range for the next pass (Supplementary Figure E.6). For example in our data set, a similarity threshold of 0.85 resulted in one exclusive group (21 strains) among the 451 strains and we deemed this PGT1. The next iterative sort identified a similarity threshold of 0.86 resulting in one exclusive group (6 strains).

The STAG algorithm categorized our dataset of 451 *C. difficile* strains containing 4,057 accessory gene clusters into 176 PGTs that comprise strain groupings ranging from one to 23 strains. We assigned MLST types to each genome using PubMLST [61] (Figure 7.4A) resulting in a total of 57 STs. Ribotype information was only available for a total of 108 strains in our dataset, limiting our direct PGT-MLST-RT comparisons to 108 genomes (Supplementary Figure E.7). Given the similar level of discriminatory power between MLST and RT and the paucity of RT

data for the public dataset, we used MLST as a baseline to compare strain grouping as a function of the number of strains evaluated (Figure 7.4D). As the number of strains considered continues to increase the resolution capabilities of MLST and STAG begin to diverge (Figure 7.4B). There is an intrinsic tradeoff for any strain-typing scheme in terms of resolution and compression; each scheme seeks to group strains as efficiently as possible (compression), but these groups must maintain meaning and distinguish strains at scale (resolution). The MLST and RT systems will result in a larger number of strains classified into fewer groups whereas the PGT maintains flexibility to establish new groups as more genetic content is considered with each additional strain used to construct the pan-genome.

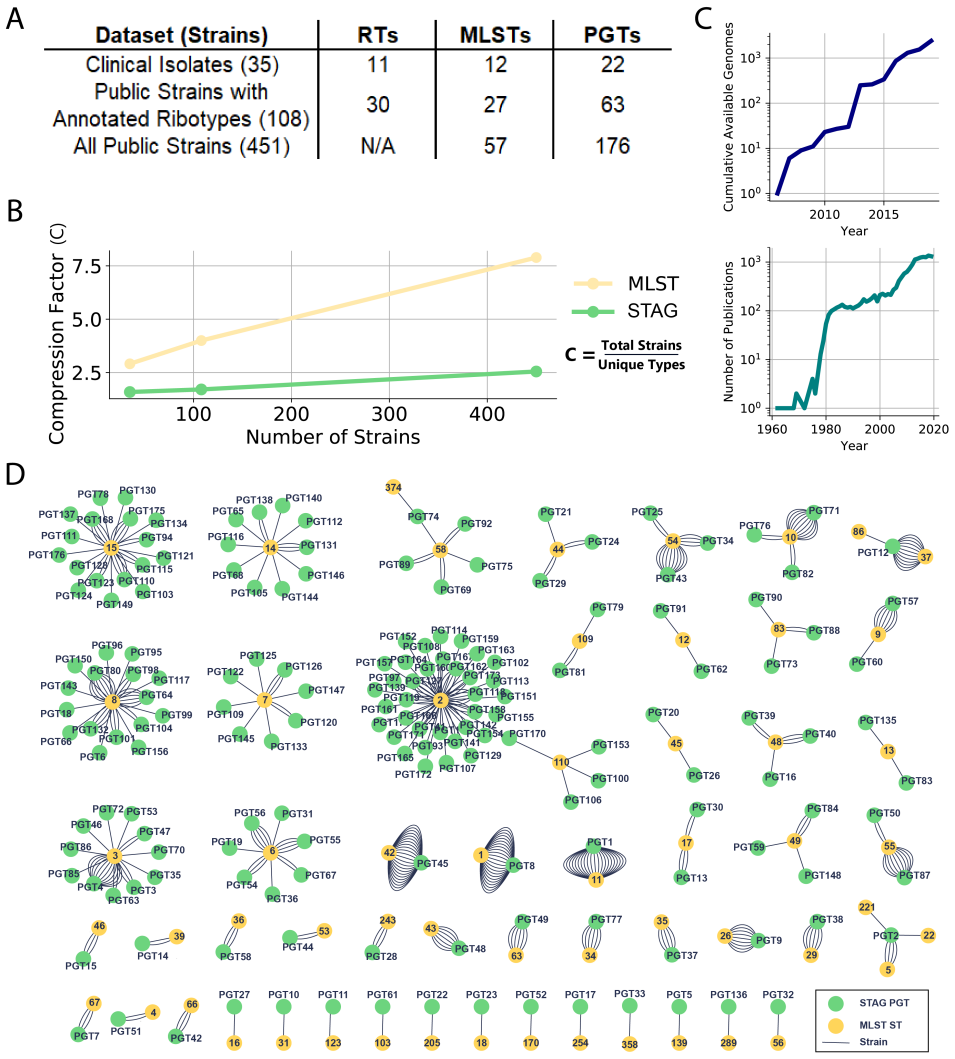

Figure 7.4: Dataset described via Ribotyping, MLST, and STAG and relative effect of dataset scale. A) Table describing the 3 levels within our dataset. Our beginning set of clinical isolates have all been ribotyped and an additional 73 public strains also had ribotyping data. MLST sequence types and STAG pan-genome types are able to be assigned for all strains. B) The compression factor as a function of the number of strains typed demonstrates that as more strains are considered strain-typing schemes like MLST do not maintain their resolution whereas the STAG scheme is comparatively invariant to the scale of strains considered. C) The importance of considering scale in terms of the growing amount of genomic data available and expanding bibliome for *C. difficile*, both of which are best visualized on a log-linear scale. These plots demonstrate that methods considering scale will only be of increasing importance. D) For each strain the two types assigned through either STAG or MLST are represented through this network where the links are each of the 451 strains studied connected to nodes of strain types for each scheme. This analysis highlights the relative number of strains of each type within the dataset as well as the certain MLST types where there is sufficient accessory gene diversity among the strains that STAG establishes numerous different PGTs.

7.3.5 STAG types exhibit an enhanced ability to explain unique metabolic profiles

We cross-referenced our STAG PGT schemes against the Biolog Phenotype Microarray profiles from the 35 isolates in our experimental dataset to determine if STAG types provided increased resolution of distinct metabolic profiles across strains. For these 35 isolates, 28 compounds exhibited differential binary growth capabilities providing a distinct binary growth vector defining the metabolic profile for each strain (Supplementary Figure E.1). The distribution of binary growth capabilities across the 35 strains resulted in 26 unique metabolic profiles, where the profile shared by the greatest number of strains (three) was defined by growth supporting carbon utilization on 3 of the 28 discriminatory compounds by the strains CDH718, CDS009, and CDS079. In turn the three strain typing schemes classified the 35 strains into 11 distinct PCR ribotypes, 12 MLST sequence types, and 22 STAG PGTs. To study the relationship between these categorical variables (metabolic profiles and strain types) we employed an asymmetric (non-linear) measure of association by calculating the uncertainty coefficient based on conditional entropy (Methods). The uncertainty coefficient indicates what fraction of information can be predicted from one variable when given the other variable. In this case we are strictly interested in evaluating how well strain type informs experimental metabolic profile, where a value of 0 would be no association and 1 would be an exact prediction. MLST and RT had uncertainty coefficients of 0.57 and 0.53 respectively, whereas PGT resulted in an uncertainty coefficient of 0.80.

While the PGTs are more informative of metabolic profiles, we note that this increase is likely a function of the difference in number of labels to describe the strains by each typing scheme. When using PGTs the 35 strains are described by 22 labels whereas the RT and MLST

describe the strains as 11 and 12 labels respectively. In an effort to evaluate the scalability of this metric we utilized our 35 strain-specific GEMs to generate draft GEMs for all 415 public strains within our dataset and generated *in silico* growth predictions as an approximation for metabolic profiles. Our approximation for growth capabilities of the 451 strains resulted in the definition of 19 in silico metabolic profiles for which the uncertainty coefficient of the strain typing schemes for MLST and PGT was calculated. Here the MLST and PGT calculated uncertainty coefficients of 0.85 and 0.92 respectively as a result in the shift of relative number of categorical variables at the larger sample size. Overall, this demonstrates that the PGT scheme is less variable as a function of dataset size as it performs more similarly when considering 35 versus 451 strains and in both cases informs a similarly sized set of categorical variables.

In addition to providing metrics that evaluate a typing schemes' ability to inform on overall metabolic profiles, examining specific metabolic capabilities illustrates the ability to interrogate functional diversity through STAG PGTs. The niche capability of RT078/ST11 strains to grow using trehalose as a carbon source has recently been associated with virulence implications in *C. difficile* infection[54]. The molecular basis for trehalose utilization in RT078/ST11 strains has been attributed to a four-gene insertion, which includes lower homology second copies of the canonical phosphotrehalase (TreA2) and repressor (TreR2) as well as genes encoding a potential trehalose specific PTS component (PtsT) and putative glycan debranching enzyme (TreX). We examined the accessory gene clusters used to establish STAG PGTs and identified a total of 12 gene clusters corresponding to this trehalose utilization operon: single gene clusters for *treX* and *treR2*, and 5 related but distinct gene clusters for both *treA2* and *ptsT* (Supplementary Figure E.8A). The single *treX* cluster along with cluster *treA2_4* and cluster *ptsT_2* are present within 16.1% (73/451) of the 451 strains which includes all of the RT078/ST11 strains (21) studied. The

single *treR2* cluster along with clusters *ptsT_4*, *ptsT_5*, *treA2_2*, and *treA2_5* are nearly ubiquitous to the overall population representing (450/451, 444/451, 445/451, 444/451, 391/451). Interestingly, sequences *treA2_3*, *ptsT_3*, and *treA2_1* are uniquely found in strain 1496.1669. Finally, the remaining *ptsT*-related gene cluster (*ptsT_1*) is specific to 8 strains classified by STAG as PGT2, wherein the strains here represent a mix of MLST ST5, ST22, and ST221 and critically this sequence is closest in similarity to the *ptsT_2* cluster including the RT078 strains (Supplementary Figure E.8B). STAG PGTs are based on iterative sequence comparisons as illustrated here, and the resulting PGTs reflect these relationships allowing for explicit identification of a large number of implicated genetic loci that otherwise would remain undetected.

7.3.6 Pan-genome types allow investigation of defining accessory gene content

In addition to providing a means of strain-typing that is less subject to a loss of resolution at increasing scale, the PGTs can be interrogated to study functions within the population that drive separation into calculated groups. The 176 distinct PGTs identified among the 451 genomes were compared for gene cluster presence/absence (Methods) and defining gene products were examined. These gene clusters are the drivers for inclusion within each PGT. The annotation information density for each defining group of clusters (presence or absence thereof) was calculated (using the number of genes within a gene cluster with annotation information divided by the total number of genes in the gene cluster and averaged for all clusters identified for a PGT) and used to prioritize gene clusters for deeper study (Table 7.1 and 7.2). Given the widespread literature on specific ribotype lineages known for being epidemic, we focused on the PGTs that contain the following clinically relevant ribotypes: RT078, RT027, RT017, RT106, and RT002.

Six clinical isolates from our original dataset are empirically classified as RT078 (CDH074,

Table 7.1: Pan-Genome Typings Containing at least one strain known to be of a hypervirulent ribotype and size of the PGT, number of gene clusters identified.

PGT	Epidemic RTs in PGT	PGT Size (Strains)	Presence GCs	Absence GCs
PGT1	RT078	21	124	110
PGT8	RT027	23	40	27
PGT12	RT017	13	62	11
PGT45	RT106	23	6	2
PGT95	RT002	1	10	0
PGT96	RT002	1	20	2
PGT98	RT002	3	12	0
PGT99	RT002	1	7	0
PGT101	RT002	3	9	0
PGT104	RT002	2	2	0
PGT156	RT002	1	12	0

Table 7.2: Pan-Genome Typings Containing at least one strain known to be of a hypervirulent ribotype and degree of available annotation information (Annotation Information Density).

PGT	Presence Cluster COG	Absence Cluster COG	Presence Genes	Absence Genes
PGT1	0.298	0.564	0.556	0.854
PGT8	0.3	0.296	0.55	0.667
PGT12	0.435	0.545	0.726	0.909
PGT45	0.167	0	0.167	0
PGT95	0.2	0	0.3	0
PGT96	0.3	0	0.1	1
PGT98	0.083	0	0.167	0
PGT99	0.143	0	0	0
PGT101	0.111	0	0.222	0
PGT104	0	0	0	0
PGT156	0.167	0	0.083	0

CDH180, CDH333, CDS009, CDS010, CDS031). These same genomes were classified within PGT1 by the STAG method described here. PGT1 contains a total of 21 strains from the 451 genomes used to define the *C. difficile* PGT scheme, and all strains within PGT1 are also classified by pubMLST as ST11. Given the nature of the sorting algorithm used to construct PGTs, the order in which PGTs arise is an indication of the degree of uniqueness of the group and this is reflected in the fact that PGT1 is defined by the average presence of 121 gene clusters and absence of 110 gene clusters in contrast to the population of strains evaluated here. The PGT1 strains

represent the most genetically distinct group out of the 176 PGTs we have defined, a distinction that aligns with previous studies characterizing the zoonotic prevalence of RT078/ST11 [62–64]. STAG PGT classification has identified specific gene clusters that may inform the emergence and virulence of RT078 strains outside of trehalose utilization discussed above. Specifically, PGT1 contains a cluster annotated as an adaptive-response sensory kinase, *sasA*. In other clinically relevant organisms, *sasA* is responsible for binding to the innate immune receptor glycoprotein DMBT1 promoting bacterial adhesion to tissue within the oral cavity [65, 66]. DMBT1 is also found in other tissues like the lung and small intestine. The presence of *sasA* positive *C. difficile* strains could provide PGT1 strains an adhesion and colonization advantage over other *C. difficile* strains. A second PGT1-specific gene cluster of interest is the sensor histidine kinase *prrB*. Previous studies indicate that *prrB* is involved in regulating anaerobic metabolism [67, 68]. Furthermore, PGT1 includes three additional gene clusters involved in the acquisition and homeostasis of zinc; textitznuA, *znuB*, and *yeiR*. Characterization of the *znuA/znuB* system in *Acinetobacter baumannii* has demonstrated roles in resistance to calprotectin-mediated chelation of zinc, which has been suggested to be a strategy to circumvent nutritional immunity [69, 70]. While these genes are present throughout the *A. baumannii* species, these gene clusters are only identified exclusively to PGT1 *C. difficile* strains. The importance of zinc acquisition is further supported by the presence of PGT1 exclusive *yeiR*, which has also been implicated in metal homeostasis in *E. coli* [71]. Finally, the presence of a tellurium resistance protein TerC is identified as one of the gene clusters driving PGT1 separation. Tellurium resistance genes have been shown to have low levels of divergence and these resistance genes are thought to be widespread among pathogenic bacteria [72, 73].

The most prevalent *C. difficile* ribotype among hospital-associated CDI in the United

States is RT027[74, 75]. Strains within RT027 are considered hypervirulent and have persisted as the dominant clone in hospital-associated infections since their emergence in the early 2000s. Our dataset includes 2 clinical isolates (CDH352, CDS041) from RT027 that are classified by MLST as ST1 and by our new pan-genome typing method as PGT8. PGT8 also contains an additional 21 publicly available genomes also classified as ST1. PGT8 is defined by 40 present and 27 absent accessory gene clusters, and several of the annotated clusters have potential implications to contribute to the hypervirulent nature of these strains. Like PGT1, PGT8 includes an additional distinct gene cluster annotated as an adaptive-response sensory kinase (*sasA*). With a clustering identity threshold of 80%, we have identified 7 of the 4,057 accessory gene clusters with this annotated function. Each cluster contains genes from a small number of strains ranging from 1.1% (5/451) to 6.2% (28/451) with certain clusters such as those identified in regard to PGT8 and PGT1 being exclusive to certain PGTs. The presence of these gene clusters, particularly in groupings which include strains known to be highly problematic, points towards the potential importance of this feature within the evolutionary trajectory of the species. PGT8 contains a gene cluster annotated as *yxdL* which has been shown to be an ABC transporter participating in a genomic structure of adjacent two-component systems and related ABC transporter, a feature associated with *B. subtilis* and *Clostridia* genomes [76, 77]. While the full function of *yxdL* remains unknown, evidence suggesting that it functions as an antibiotic efflux pump is supported by homology to *salX*, which confers salivaricin resistance in *Streptococcus salivarius* [76]. Another gene cluster implicated within PGT8 is annotated as *bceB*, a bacitracin export permease protein. Furthermore the bce system is paralogous to the yxd system and a component of bacitracin resistance [78]. From a metabolic standpoint PGT8 also contains a gene cluster indicated as *potA*, a spermidine/putrescine transport system that has been studied in *E. coli*

[79]. Interestingly, spermidine biosynthesis pathway genes and transporter components, including *potA*, have been shown to be up-regulated during temperature and alkali stress in *C. difficile* [80]. PGT8 clusters also include the presence of thymidylate synthase and phosphomethylpyrimidine synthase suggesting isozymes within the species for these functions. Finally, it is worth noting that PGT8 contains IS3 and IS1595 family transposases indicating potentially consistent mobile elements among the strains.

RT017 is a unique virulent lineage because it is toxinA negative/toxinB positive [81, 82]. PGT12 in total contains 13 strains, three of which are known to be RT017 (M68, 1141436.4, 1151438.4). All strains within PGT12 are typed by MLST as either MLST37 or MLST86 in agreement with previous studies of this lineage[83]. PGT12 is defined by 62 present and 11 absent gene clusters that contain a high degree of annotation information predominated by gene transcriptional regulator annotations. Of note is a cluster annotated as N-acetylmuramoyl-L-alanine amidase which is associated with bacteriophage endolysin activity [84, 85]. We also analyzed within PGT12, which gene clusters distinguish the MLST37 from MLST86 strains and were able to identify 75 clusters that contrasted each other within the PGT12, which the majority of remain poorly annotated, but does include peptidoglycan acetyltransferase and membrane protein specific to strains of MLST37 and a proline transporter specific to MLST86.

RT106 reflects the most prevalent community-acquired ribotype according to CDC surveillance and the 2nd most healthcare acquired ribotype to date[86]. We had three known RT106 strains within our dataset (CDH054, CDH220, CDS057) and all of these strains were grouped into PGT45. There are 23 strains, all MLST42, within PGT45 that are defined by 6 present clusters and 2 absent clusters, and very limited annotation information overall. Interestingly, PGT45 also contains CDH718 which is known to be RT014 and 5 of the public strains

annotated as RT_SW11. The lone gene cluster with annotation information is annotated as "IS110 family transposase ISFnu3". This uniquely present mobile genetic element within the strains of a known problematic ribotype could reflect the acquisition of an adaptive trait.

The last clinically relevant ribotype of interest was RT002, another highly healthcare-acquired ribotype for which there were 8 total strains in our dataset (CDS064, CDS065, 1151326.4, 1151354.4, 1151373.4, 1151375.4, 1151403.4, 1151418.4). These RT002 strains, while only classified into ST8 by MLST, were classified into 6 PGTs: PGT96 (1/8), PGT98 (2/8), PGT99 (1/8), PGT101 (1/8), PGT104 (1/8), and PGT156 (1/8). Although fraught with a paucity of annotation information, we were able to identify notable functional characteristics within this set of PGTs. PGT95 was defined by a cluster annotated as *ydpB*, which is a component of the *ypdA/ypdB* histidine kinase/response regulator pair. Previous studies within *E. coli* demonstrated that this system responds to extracellular pyruvate and is indicated in growth phase-dependent regulation in response to the availability of carbon sources [87, 88]. PGT96 was partially defined by two absent clusters that encode penicillinase repressors known to play a key role in the regulation of penicillinase synthesis within gram-positive bacteria.[89] The absence of the repressor in this strain could indicate the constitutive expression of the penicillinase synthesis genes and increased antibiotic resistance. Lastly, within PGT156 the gene encoding cell wall-binding protein cwp26 is uniquely present. *C. difficile* is known to produce a number of surface proteins that comprise the S-layer and these proteins are suspected to have roles in pathogenesis [90, 91] and the cwp26 contains a putative functional domain of PepSY, which is predicted to have protease inhibition function.

If the pan-genome is separated into its constituent functional annotations (Figure 7.3CD) the strains can be classified using STAG on specific functional subsections (Supplementary Text

S1). Interestingly when RT002 strains are typed according to metabolically relevant gene clusters all strains are grouped into one type of 36 strains. Cluster significance of this metabolically relevant grouping shows that there are seven clusters absent within these strains that are present within 77% of the overall population, and another 2 clusters absent that are present within 58% of the population. Analyzing the functional annotations available for these clusters demonstrates that 5 of these clusters correspond to various genes within the *yxe* operon, which has been characterized in the related species *Bacillus subtilis* [92–94]. The implicated genes within the operon have been shown to be primary transporters of the ATP binding cassette variety (ABC) for polar amino acid uptake and in a more recent study as key pieces of a disposal route for S-(2-succino)cysteine (2SC). 2SC is a product of fumarate-mediated succination of thiols [95], a process implicated to increase in certain tumors, diabetes, and obesity. The presence of this compound could be used as a biomarker indicating higher levels of cellular aerobic respiration that may result in tumorigenesis, diabetes, and/or obesity[96–98]. The absence of this operon within the metabolically clustered RT002 strains may lead to the inability of RT002 strains to use 2SC as a sulfur source, resulting in greater concentrations of 2SC in the gut after invasion of an RT002 strain. Of the remaining absent gene clusters three are annotated as C4-dicarboxylate transport protein [99], phospho-beta-D-glucosidase *bglH* [100], and l-cystine transport permease protein and one cluster with no valuable annotation information. The C4-dicarboxylate transport protein encoding gene has been shown to be a participant of the sigma G regulon in sporulation and its product detected in *C. difficile* spores.

7.4 Discussion

In this study, we perform a functional analysis of the *C. difficile* pan-genome in an effort to increase understanding of strain-specific traits in terms of both genotype and phenotype. Taking a systems biology approach enabled us to identify and contextualize important genetic and phenotypic features within the vast diversity of this species. Motivated by the importance of specific carbohydrate and bile acid metabolism in *C. difficile* pathogenesis [57, 101–103], we metabolically profiled 35 clinical isolates and investigated their diverse capabilities. The wide array of growth dynamics exhibited from our high-throughput screening necessitated sophisticated data analysis, which was facilitated by the use of gaussian process regression models. These two techniques demonstrated through variable growth modalities that catabolic capabilities were diverse at a strain-specific level including differences across strains of the same PCR ribotype and MLST sequence type. Following the identification of unique carbon source utilization profiles, strain-specific genome-scale models of metabolism were generated for each isolate to bridge the genotype to observed phenotypic diversity and infer potential mechanistic insight. The in silico simulations recapitulated the majority (76%) of growth phenotypes. However, there were a high number of false positive error mode predictions, which indicated that the models of metabolism, which are predictors of all theoretically possible growth capabilities based on enzymatic coding gene content, were lacking the biological context concerning transcriptional regulation and/or enzyme efficiency that restrict capabilities in vitro [104].

To robustly explore all the genetic diversity outside of the metabolic network, we constructed the pan-genome of *C. difficile* with the inclusion of an additional 416 public genomes. Characterizing the pan-genome demonstrated different conservation levels across various functional categories. We used the accessory genome profile to type our group of 451 strains through

the novel development of the STAG algorithm. We were able to identify gene clusters that strongly contributed to unique groupings of strains based on their contrasting presence and absence from the overall population. The functional pan-genomics approach brought to the surface traits ranging from specific transporters, sensory responses, two component systems, to cell wall proteins across the clusters driving separation of PGTs containing known epidemic lineages. An especially valuable aspect of the approach is identification of a large and diverse number of new genetic loci that differentiate strains. These loci present critical candidates for further characterization and improvement of annotation to increase understanding of pathogenesis at the species level.

Overall, the results presented here suggest the importance of a genomics driven approach to understand *C. difficile* diversity and identification of the evolutionary events leading to propagation of epidemic lineages. Trait acquisition has been demonstrated across functional categories and most pressing is the vast amount of genetic content that remains uncharacterized. The high percentage (74.5%) of implicated present genes with poor to no annotation information within the gene clusters driving separation of PGTs demonstrates that overall characterization of genes lacking experimental evidence of function (the "y-ome") [105] for *C. difficile* remains high. Unsurprisingly, these clusters exhibit the highest degree of openness within the subdivisions of the pan-genome and likely these clusters contain genes that are critical factors in the evolutionary trajectory and history of *C. difficile*. Our exploration of total gene content has suggested that an investigation into the transcriptional regulatory network of *C. difficile* would prove informative. The processes involved and related to regulation appear to be critical in differentiating strains and an accurate description of the transcriptome in presumed physiological conditions during infection would provide a crucial systems-level explanation of cellular response. Use of machine

learning methods on high-quality expression profiles has been shown to provide such a window into understanding transcriptional regulation in *E. coli* and *S. aureus* [106, 107] and with proper datasets could be applied to *C. difficile*.

The insights into the accessory genome and its specific components to groups of strains presented here has added to the overall understanding of *C. difficile* and provided a means for bringing the important factor of genetic diversity to the forefront. Our use of the accessory genome through STAG demonstrates the ability to extract knowledge from big data through a method less subject to resolution loss at scale when compared to traditional approaches such as PCR ribotyping and MLST. The STAG method presented has advantages in maintaining flexibility with the scale of strains studied, reliance on solely WGS data, ability to identify functional differences across PGTs, and the illumination of new genetic loci with discriminatory power. In any strain typing scheme there will be a tradeoff between compression and resolution of the resulting groups in that each scheme strives to establish meaningful groups that capture the relationship among strains. Given the continued growth of genome sequences available for most bacterial species, methods that leverage this data to identify key genetic features in relation to populations will be important to the future of global epidemiology. Future endeavors in characterization in concert with data analytics will enhance the scientific knowledge of the *C. difficile* species commensurate with the promise of omics big data.

7.5 Methods

7.5.1 Phenotypic Profiling by Biolog

Strains were cultured in BHI medium (Difco) supplemented with 0.5% (w/v) yeast extract (Fisher Scientific) overnight (16 hours) in an anaerobic chamber (5% hydrogen, 90% nitrogen, 5% carbon dioxide). 1 ml of overnight culture was diluted into 10 ml of defined minimal media with previously described composition [108] and 100 µl was added to each well of Biolog Phenotypic Microarray plates (PM1 and PM2). Growth assays were performed under anaerobic conditions with optical density at 620 nm read every 10 minutes following 5 seconds of shaking over a period of 16 hours.

7.5.2 Gaussian Process Regression Models of Growth

Correlation between experimental replicates were evaluated and replicates passing this quality check (Pearson R>0.7) were pooled for the following analysis. One of the advantages to using gaussian process (GP) regression is the ability to pool biological replicates and makes this approach particularly suited to high-throughput screens. GP to infer microbial growth parameters was conducted using the AMiGA [109] program through the default settings presented by AMiGA. For each strain the pooled quality replicates were log transformed and negative control subtracted at each time point. A GP regression model was then fit for each pooled time course and growth parameters inferred resulting in identification of the growth parameters for each strain on each of the 95 carbon sources.

7.5.3 Whole Genome Sequencing

Cryofrozen isolates of each *C. difficile* strain were incubated on Brain Heart Infusion (BHI) agar under anaerobic conditions for 24-48 h. Genomic DNA was extracted using the MasterPure Complete DNA RNA Purification kit (Lucigen, MC85200) and libraries of fragmented genomic DNA were prepared using NEXTflex Rapid DNA-Seq Kit (Bioo Scientific, NOVA-5149-02). Paired-end reads (2 x 150 bp reads) were generated on the MiSeq platform (Illumina, San Diego, CA, USA) using the Illumina MiSeq Reagent Kit v2 (MS-102-2002) and PhiX Control Kit v3 (FC-110-3001). WGS reads have been submitted to NCBI as BioProject PRJNA472399.

7.5.4 Sensitivity Analysis of Growth Dynamics Parameters

Potential thresholds for each growth parameter (AUC or K) were evaluated on the range from minimum to maximum parameter value across all strains. Each parameter was evaluated separately by binarizing experimental data using each potential threshold, while holding the complementary parameter threshold constant. This led to identification of greater than 1.25 AUC and greater than .3 for K to jointly define growth calls of the experimental data.

7.5.5 Constraint-based modeling flux balance analysis

Constraints-based analyses were conducted using the COBRApy toolbox. For the in silico growth simulation of sole carbon source utilization the minimal media [110] was used and glucose was removed and all other carbon source exchange reactions were opened in an iterative fashion to evaluate if growth was possible [31, 36, 44, 49, 51]. Growth versus no growth was determined through flux balance analysis in each condition, optimizing for the biomass function. Within these simulations we consider biomass objective flux of greater than zero designated

carbon sources that supported growth.

7.5.6 Strain-Specific Model Creation

A standard strain-specific model generation protocol [48] was followed to generate draft strain-specific models for the dataset of 35 isolates as well as 415 publicly available strains. In the case of the 35 isolates the outputs of these 35 isolates were further curated based on false negative predictions identified through comparison to our experimental dataset [111]. This manual curation involved identification of reactions and genes through literature curation as well as homology with other related species. Following the curation of the 35 strain-specific models, the isolate closest in terms of phylogeny to each public strain was used as the base model for generation of draft models for the public strains.

7.5.7 Pan-Genome Construction and Analyses

A total of 1,246 whole-genome sequences of *C. difficile* were downloaded from the PATRIC database [112] on August 25, 2019. To filter for high-quality genomes a cutoff of assemblies composed of 100 or fewer contigs was applied. Furthermore, an MLST analysis of the genomes was performed using MLST [61]. All genomes that could not be assigned to an MLST type or species were also filtered out. This led to a final set of 415 genome sequences for downstream analysis. Sequence homology was used to cluster genes into gene families using CD-Hit [113]. Clustering was performed with 0.8 threshold and word length of 5. These gene families were then used to designate core and pan genes by identifying the gene families found in less than 1% of the total 451 strains. Biserial correlations between all gene clusters and measured phenotypes were calculated using the pointbiserial function in the scipy stats package. The Cramer's V statistic

for categorical-to-categorical association was used to evaluate associations between pan-genome clusters and ST types. This calculation was implemented in python and uses the correction from Bergsma and Wicher [114].

7.5.8 Phylogenomic Analysis

Phylogenomic analysis was performed using GToTree [115] on the clinical isolate dataset along with 415 public strains. Prokka FASTA files were used as inputs to GToTree analysis and the resulting tree was visualized using the Interactive Tree of Life web-based tool [116, 117]. This tree was compared with a dendrogram constructed using a binary representation of the accessory genome (451x4,057 maxtrix where 1=gene cluster present, 0=gene cluster absent). The dendrogram was constructed using jaccard distances and the complete clustering method. The two trees were compared (including calculation of correlation and entanglement) using the dendextend package in R [118]. Clustering methods were compared to those of cgMLST (extracted from enterobase[29]) and KMERs obtained using the SKA package [119].

7.5.9 Using Jaccard Similarity to Establish Strain Groups

Using the constructed pan-genome the accessory gene clusters (those that are not core nor unique) were identified. These 4,057 gene clusters were used to define a vector of presence/absence for each strain in terms of whether or not that strain had genes present within the gene cluster. The Jaccard Similarity coefficient was calculated between each of these 451 binary vectors of length 4,057 and used to establish a symmetric 451x451 matrix of the similarity coefficients. An iterative sorting workflow was developed that parses through this matrix and establishes pan-genome typings. A range of potential similarity thresholds is established and the set of strains

meeting the current Jaccard similarity matrix threshold for each single strain are identified. From this information exclusive groups (i.e. at the threshold, every strain similar to the current strain is also sufficiently similar to every other strain similar to a given strain) are identified on a per threshold basis. Out of these exclusive groups the compression factor (Number of Strains in Exclusive Groups/ Number of Exclusive Groups) is calculated and the threshold maximizing this compression factor is selected for this iteration over the Jaccard Similarity matrix. The exclusive groups are established as PGTs and corresponding strains are filtered from the Jaccard Similarity Matrix for the next sorting pass. Additionally, the compression maximizing threshold is the starting point of similarity threshold range. This process is repeated until the similarity threshold has been incremented to only identify single strains into PGTs.

7.5.10 Identification of Gene Clusters Driving PGT Separation

For comparison of each PGT the binary presence/absence vectors of each strain within the group were averaged to calculate the representative gene cluster portfolio for each PGT. The PGT-specific mean and all strain population mean were compared to identify divergent clusters. We consider a gene cluster present in less than 20% of the population to be predominantly absent in the overall population and a gene cluster present in greater than 90% of the population to be predominantly present in the overall population. The absolute value of the difference between mean presence in population versus mean presence in a specific PGT was calculated for each cluster and used to designate as either predominantly present or absent. In each case if the absolute value of difference was greater than 0.95 the cluster was identified as being present or absent the given PGT in contrast to the population. For example if a gene cluster was determined to only be present within 15% of the overall strains, but was present in 100% of the strains of

a given PGT, this cluster would be identified as a driving cluster in terms of gene presence for that PGT. For each cluster identified in this way the functional annotations at both the gene cluster level (in terms of COG) and gene annotation level (in terms of function) were evaluated. The information density (ratio of clusters with annotation information to total clusters) was calculated for each PGTs defining clusters in terms of both presence and absence. To ensure that information was not simply more available for PGTs with a greater number of clusters identified the size of PGT in terms of number of strains and number of clusters driving separation of the PGT was evaluated (Number of Presence Clusters R=.145 , Number of Absent Clusters R=.372). Additionally, correlation between PGT size and information density was evaluated (Number of Presence Clusters R=.249 , Number of Absent Clusters R=.089).

Acknowledgements

Chapter 7, in part, is currently being prepared for submission for publication: **Norsigian CJ**, Danhof HA, Brand CK, Midani FS, Broddrick JT, Savidge TC, Britton RA, Palsson BO, Spinler JK, Monk JM. "Systems biology approach to functionally assess the *Clostridioides difficile* pan-genome reveals genetic diversity with discriminatory power." The dissertation author is the primary author.

7.6 References

1. For Disease Control, C., Prevention, *et al.* Biggest Threats and Data 2019 AR Threats Report (2019).
2. Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. & Kuijper, E. J. Clostridium difficile infection. en. *Nat Rev Dis Primers* **2,** 16020 (Apr. 2016).

3. Martin, J. S. H., Monaghan, T. M. & Wilcox, M. H. Clostridium difficile infection: epidemiology, diagnosis and understanding transmission. en. *Nat. Rev. Gastroenterol. Hepatol.* **13,** 206–216 (Apr. 2016).

4. Rupnik, M., Wilcox, M. H. & Gerding, D. N. Clostridium difficile infection: new developments in epidemiology and pathogenesis. en. *Nat. Rev. Microbiol.* **7,** 526–536 (July 2009).

5. Kuehne, S. A., Cartman, S. T., Heap, J. T., Kelly, M. L., Cockayne, A. & Minton, N. P. The role of toxin A and toxin B in Clostridium difficile infection. en. *Nature* **467,** 711–713 (Oct. 2010).

6. Voth, D. E. & Ballard, J. D. Clostridium difficile toxins: mechanism of action and role in disease. en. *Clin. Microbiol. Rev.* **18,** 247–263 (Apr. 2005).

7. Stevenson, E., Minton, N. P. & Kuehne, S. A. The role of flagella in Clostridium difficile pathogenicity. en. *Trends Microbiol.* **23,** 275–282 (May 2015).

8. Cheng, V. C. C., Yam, W. C., Lam, O. T. C., Tsang, J. L. Y., Tse, E. Y. F., Siu, G. K. H., Chan, J. F. W., Tse, H., To, K. K. W., Tai, J. W. M., Ho, P. L. & Yuen, K. Y. Clostridium difficile isolates with increased sporulation: emergence of PCR ribotype 002 in Hong Kong. en. *Eur. J. Clin. Microbiol. Infect. Dis.* **30,** 1371–1381 (Nov. 2011).

9. Knight, D. R., Elliott, B., Chang, B. J., Perkins, T. T. & Riley, T. V. Diversity and Evolution in the Genome of Clostridium difficile. en. *Clin. Microbiol. Rev.* **28,** 721–741 (July 2015).

10. Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J., Frost, E. & McDonald, L. C. Toxin production by an emerging strain of Clostridium difficile associated with outbreaks of severe disease in North America and Europe. en. *Lancet* **366,** 1079–1084 (2005).

11. McDonald, L. C., Killgore, G. E., Thompson, A., Owens, R. C., Kazakova, S. V., Sambol, S. P., Johnson, S. & Gerding, D. N. An Epidemic, Toxin Gene–Variant Strain of Clostridium difficile. *N. Engl. J. Med.* **353,** 2433–2441 (Dec. 2005).

12. Goorhuis, A., Bakker, D., Corver, J., Debast, S. B., Harmanus, C., Notermans, D. W., Bergwerff, A. A., Dekker, F. W. & Kuijper, E. J. Emergence of Clostridium difficile infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. *Clin. Infect. Dis.* **47,** 1162–1170 (2008).

13. Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., Thomson, N. R., Roberts, A. P., Cerdeño-Tárraga, A. M., Wang, H., Holden, M. T. G., Wright, A., Churcher, C., Quail, M. A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., Moule, S., Mungall, K., Price, C., Rabbinowitsch, E., Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B. & Parkhill, J. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat. Genet.* **38,** 779 (June 2006).

14. Brouwer, M. S. M., Roberts, A. P., Hussain, H., Williams, R. J., Allan, E. & Mullany, P. Horizontal gene transfer converts non-toxigenic Clostridium difficile strains into toxin producers. en. *Nat. Commun.* **4,** 2601 (2013).

15. Sadeghifard, N., Gürtler, V., Beer, M. & Seviour, R. J. The mosaic nature of intergenic 16S-23S rRNA spacer regions suggests rRNA operon copy number variation in Clostridium difficile strains. en. *Appl. Environ. Microbiol.* **72,** 7311–7323 (Nov. 2006).

16. Gürtler, V. Typing of Clostridium difficile strains by PCR-amplification of variable length 16S-23S rDNA spacer regions. en. *J. Gen. Microbiol.* **139,** 3089–3097 (Dec. 1993).

17. Indra, A., Huhulescu, S., Schneeweis, M., Hasenberger, P., Kernbichler, S., Fiedler, A., Wewalka, G., Allerberger, F. & Kuijper, E. J. Characterization of Clostridium difficile isolates using capillary gel electrophoresis-based PCR ribotyping. en. *J. Med. Microbiol.* **57,** 1377–1382 (Nov. 2008).

18. Janezic, S. Direct PCR-Ribotyping of Clostridium difficile. en. *Methods Mol. Biol.* **1476,** 15–21 (2016).

19. Bauer, M. P., Notermans, D. W., van Benthem, B. H. B., Brazier, J. S., Wilcox, M. H., Rupnik, M., Monnet, D. L., van Dissel, J. T., Kuijper, E. J. & ECDIS Study Group. Clostridium difficile infection in Europe: a hospital-based survey. en. *Lancet* **377,** 63–73 (Jan. 2011).

20. Freeman, J., Vernon, J., Morris, K., Nicholson, S., Todhunter, S., Longshaw, C., Wilcox, M. H. & Pan-European Longitudinal Surveillance of Antibiotic Resistance among Prevalent Clostridium difficile Ribotypes' Study Group. Pan-European longitudinal surveillance of antibiotic resistance among prevalent Clostridium difficile ribotypes. en. *Clin. Microbiol. Infect.* **21,** 248.e9–248.e16 (Mar. 2015).

21. Griffiths, D., Fawley, W., Kachrimanidou, M., Bowden, R., Crook, D. W., Fung, R., Golubchik, T., Harding, R. M., Jeffery, K. J. M., Jolley, K. A., Kirton, R., Peto, T. E., Rees, G., Stoesser, N., Vaughan, A., Walker, A. S., Young, B. C., Wilcox, M. & Dingle, K. E. Multilocus sequence typing of Clostridium difficile. en. *J. Clin. Microbiol.* **48,** 770–778 (Mar. 2010).

22. Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M. & Spratt, B. G. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. en. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 3140–3145 (Mar. 1998).

23. Lemée, L. & Pons, J.-L. Multilocus sequence typing for Clostridium difficile. en. *Methods Mol. Biol.* **646,** 77–90 (2010).

24. Huber, C. A., Foster, N. F., Riley, T. V. & Paterson, D. L. Challenges for standardization of Clostridium difficile typing methods. en. *J. Clin. Microbiol.* **51,** 2810–2814 (Sept. 2013).

25. Chain, P. S. G., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C., Buhay, C., Cole, J. R., Ding, Y., Dugan, S., Field, D., Garrity, G. M., Gibbs, R., Graves, T., Han, C. S., Harrison, S. H., Highlander, S., Hugenholtz, P., Khouri, H. M., Kodira, C. D., Kolker, E., Kyrpides, N. C., Lang, D., Lapidus, A., Malfatti, S. A., Markowitz, V., Metha, T., Nelson, K. E., Parkhill, J., Pitluck, S., Qin, X., Read, T. D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R. L., Sutton, G., Thomson, N. R., Tiedje, J. M., Weinstock, G., Wollam, A., Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium & Detter, J. C. Genomics. Genome project standards in a new era of sequencing. en. *Science* **326,** 236–237 (Oct. 2009).

26. Quainoo, S., Coolen, J. P. M., van Hijum, S. A. F. T., Huynen, M. A., Melchers, W. J. G., van Schaik, W. & Wertheim, H. F. L. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. en. *Clin. Microbiol. Rev.* **30,** 1015–1063 (Oct. 2017).

27. Janezic, S. & Rupnik, M. Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for Clostridioides difficile. en. *Front Public Health* **7,** 309 (Oct. 2019).

28. Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A. & McCarthy, N. D. MLST revisited: the gene-by-gene approach to bacterial genomics. en. *Nat. Rev. Microbiol.* **11,** 728–736 (Oct. 2013).

29. Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., Agama Study Group & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. en. *Genome Res.* **30,** 138–152 (Jan. 2020).

30. Bletz, S., Janezic, S., Harmsen, D., Rupnik, M. & Mellmann, A. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of Clostridium difficile. en. *J. Clin. Microbiol.* **56** (June 2018).

31. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nat. Rev. Genet.* **15,** 107–120 (Feb. 2014).

32. Ahmed, N., Dobrindt, U., Hacker, J. & Hasnain, S. E. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. en. *Nat. Rev. Microbiol.* **6,** 387–394 (May 2008).

33. Joyce, E. A., Chan, K., Salama, N. R. & Falkow, S. Redefining bacterial populations: a post-genomic reformation. en. *Nat. Rev. Genet.* **3,** 462–473 (June 2002).

34. Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Scott Durkin, A., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Ros, I. M. y., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N.,

Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13950–13955 (Sept. 2005).

35. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. en. *Curr. Opin. Genet. Dev.* **15,** 589–594 (Dec. 2005).

36. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20338–20343 (Dec. 2013).

37. 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. en. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3801–9 (June 2016).

38. Hoonmo L. Koo, Jennifer K. Spinler, Robert L. ... Atmar Tor C. Savidge, Herbert L. DuPont. Active Surveillance and Contact Isolation of Asymptomatic Clostridioides (Clostridium) Difficile Patients to Prevent C. difficile Infection. *Submitted* (2020).

39. Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & van 't Riet, K. Modeling of the bacterial growth curve. en. *Appl. Environ. Microbiol.* **56,** 1875–1881 (June 1990).

40. Swain, P. S., Stevenson, K., Leary, A., Montano-Gutierrez, L. F., Clark, I. B. N., Vogel, J. & Pilizota, T. Inferring time derivatives including cell growth rates using Gaussian processes. en. *Nat. Commun.* **7,** 13766 (Dec. 2016).

41. Tonner, P. D., Darnell, C. L., Engelhardt, B. E. & Schmid, A. K. Detecting differential growth of microbial populations with Gaussian process regression. en. *Genome Res.* **27,** 320–333 (Feb. 2017).

42. Norsigian, C. J., Kavvas, E., Seif, Y., Palsson, B. O. & Monk, J. M. iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of Acinetobacter baumannii AYE. en. *Front. Genet.* **9,** 121 (Apr. 2018).

43. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nat. Commun.* **9,** 3771 (Sept. 2018).

44. Norsigian, C. J., Attia, H., Szubin, R., Yassin, A. S., Palsson, B. Ø., Aziz, R. K. & Monk, J. M. Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase-Producing Multidrug-Resistant Klebsiella pneumoniae Clinical Isolates. en. *Front. Cell. Infect. Microbiol.* **9,** 161 (May 2019).

45. Norsigian, C. J., Fang, X., Palsson, B. O. & Monk, J. M. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds Tettelin, H. & Medini, D.) (Springer, Cham (CH), May 2020).

46. Norsigian, C. J., Pusarla, N., McConn, J. L., Yurkovich, J. T., Dräger, A., Palsson, B. O. & King, Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. en. *Nucleic Acids Res.* (Nov. 2019).

47. Norsigian, C. J., Danhof, H. A., Brand, C. K., Oezguen, N., Midani, F. S., Palsson, B. O., Savidge, T. C., Britton, R. A., Spinler, J. K. & Monk, J. M. Systems biology analysis of the Clostridioides difficile core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence. en. *NPJ Syst Biol Appl* **6,** 31 (Oct. 2020).

48. Norsigian, C. J., Fang, X., Seif, Y., Monk, J. M. & Palsson, B. O. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. en. *Nat. Protoc.* **15,** 1–14 (Jan. 2020).

49. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–248 (Mar. 2010).

50. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. en. *Nat. Rev. Microbiol.* **2,** 886–897 (Nov. 2004).

51. 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).

52. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C. & Bork, P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. en. *Nucleic Acids Res.* **47,** D309–D314 (Jan. 2019).

53. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nat. Biotechnol.* **35,** 904–908 (Oct. 2017).

54. Collins, J., Robinson, C., Danhof, H., Knetsch, C. W., van Leeuwen, H. C., Lawley, T. D., Auchtung, J. M. & Britton, R. A. Dietary trehalose enhances virulence of epidemic Clostridium difficile. en. *Nature* **553,** 291–294 (Jan. 2018).

55. Bartell, J. A., Blazier, A. S., Yen, P., Thøgersen, J. C., Jelsbak, L., Goldberg, J. B. & Papin, J. A. Reconstruction of the metabolic network of Pseudomonas aeruginosa to interrogate virulence factor synthesis. en. *Nat. Commun.* **8,** 14631 (Mar. 2017).

56. Folkvardsen, D. B., Norman, A., Andersen, Å. B., Rasmussen, E. M., Lillebaek, T. & Jelsbak, L. A Major Mycobacterium tuberculosis outbreak caused by one specific genotype in a low-incidence country: Exploring gene profile virulence explanations. en. *Sci. Rep.* **8,** 11869 (Aug. 2018).

57. Lewis, B. B., Carter, R. A., Ling, L., Leiner, I., Taur, Y., Kamboj, M., Dubberke, E. R., Xavier, J. & Pamer, E. G. Pathogenicity Locus, Core Genome, and Accessory Gene Contributions to Clostridium difficile Virulence. en. *MBio* **8** (Aug. 2017).

58. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nat. Commun.* **9,** 4306 (Oct. 2018).

59. Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial Resistance Prediction in PATRIC and RAST. en. *Sci. Rep.* **6,** 27930 (June 2016).

60. Real, R. & Vargas, J. M. The Probabilistic Basis of Jaccard's Index of Similarity. *Syst. Biol.* **45,** 380–385 (1996).

61. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. en. *Wellcome Open Res* **3,** 124 (Sept. 2018).

62. Knetsch, C. W., Kumar, N., Forster, S. C., Connor, T. R., Browne, H. P., Harmanus, C., Sanders, I. M., Harris, S. R., Turner, L., Morris, T., Perry, M., Miyajima, F., Roberts, P., Pirmohamed, M., Songer, J. G., Weese, J. S., Indra, A., Corver, J., Rupnik, M., Wren, B. W., Riley, T. V., Kuijper, E. J. & Lawley, T. D. Zoonotic Transfer of Clostridium difficile Harboring Antimicrobial Resistance between Farm Animals and Humans. en. *J. Clin. Microbiol.* **56** (Mar. 2018).

63. Martın-Burriel, I., Andrés-Lasheras, S., Harders, F., Mainar-Jaime, R. C., Ranera, B., Zaragoza, P., Falceto, V., Bolea, Y., Kuijper, E., Bolea, R., Bossers, A. & Chirino-Trejo, M. Molecular analysis of three Clostridium difficile strain genomes isolated from pig farm-related samples. en. *Anaerobe* **48,** 224–231 (Dec. 2017).

64. Bakker, D., Corver, J., Harmanus, C., Goorhuis, A., Keessen, E. C., Fawley, W. N., Wilcox, M. H. & Kuijper, E. J. Relatedness of human and animal Clostridium difficile PCR ribotype 078 isolates determined on the basis of multilocus variable-number tandem-repeat analysis and tetracycline resistance. *J. Clin. Microbiol.* **48,** 3744–3749 (2010).

65. Kukita, K., Kawada-Matsuo, M., Oho, T., Nagatomo, M., Oogai, Y., Hashimoto, M., Suda, Y., Tanaka, T. & Komatsuzawa, H. Staphylococcus aureus SasA is responsible for binding to the salivary agglutinin gp340, derived from human saliva. en. *Infect. Immun.* **81,** 1870–1879 (June 2013).

66. Polley, S., Louzada, S., Forni, D., Sironi, M., Balaskas, T., Hains, D. S., Yang, F. & Hollox, E. J. Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 5105–5110 (Apr. 2015).

67. Haydel, S. E., Malhotra, V., Cornelison, G. L. & Clark-Curtiss, J. E. The prrAB two-component system is essential for Mycobacterium tuberculosis viability and is induced under nitrogen-limiting conditions. en. *J. Bacteriol.* **194,** 354–361 (Jan. 2012).

68. Eraso, J. M. & Kaplan, S. Complex regulatory activities associated with the histidine kinase PrrB in expression of photosynthesis genes in Rhodobacter sphaeroides 2.4.1. en. *J. Bacteriol.* **178,** 7037–7046 (Dec. 1996).

69. Hesse, L. E., Lonergan, Z. R., Beavers, W. N. & Skaar, E. P. The Acinetobacter baumannii Znu System Overcomes Host-Imposed Nutrient Zinc Limitation. en. *Infect. Immun.* **87** (Dec. 2019).

70. Hood, M. I., Mortensen, B. L., Moore, J. L., Zhang, Y., Kehl-Fie, T. E., Sugitani, N., Chazin, W. J., Caprioli, R. M. & Skaar, E. P. Identification of an Acinetobacter baumannii zinc acquisition system that facilitates resistance to calprotectin-mediated zinc sequestration. en. *PLoS Pathog.* **8,** e1003068 (Dec. 2012).

71. Blaby-Haas, C. E., Flood, J. A., Crécy-Lagard, V. d. & Zamble, D. B. YeiR: a metal-binding GTPase from Escherichia coli involved in metal homeostasis. en. *Metallomics* **4,** 488–497 (May 2012).

72. Turkovicova, L., Smidak, R., Jung, G., Turna, J., Lubec, G. & Aradska, J. Proteomic analysis of the TerC interactome: Novel links to tellurite resistance and pathogenicity. en. *J. Proteomics* **136,** 167–173 (Mar. 2016).

73. Janvilisri, T., Scaria, J., Thompson, A. D., Nicholson, A., Limbago, B. M., Arroyo, L. G., Songer, J. G., Gröhn, Y. T. & Chang, Y.-F. Microarray identification of Clostridium difficile core components and divergent regions associated with host origin. en. *J. Bacteriol.* **191,** 3881–3891 (June 2009).

74. Valiente, E., Cairns, M. D. & Wren, B. W. The Clostridium difficile PCR ribotype 027 lineage: a pathogen on the move. en. *Clin. Microbiol. Infect.* **20,** 396–404 (May 2014).

75. *2017 Annual Report for the Emerging Infections Program for Clostridioides difficile Infection — Emerging Infections Program — HAI — CDC* `https://www.cdc.gov/hai/eip/Annual-CDI-Report-2017.html`. Accessed: 2020-6-30. June 2020.

76. Joseph, P., Guiseppi, A., Sorokin, A. & Denizot, F. Characterization of the Bacillus subtilis YxdJ response regulator as the inducer of expression for the cognate ABC transporter YxdLM. en. *Microbiology* **150,** 2609–2617 (Aug. 2004).

77. Joseph, P., Fichant, G., Quentin, Y. & Denizot, F. Regulatory relationship of two-component and ABC transport systems and clustering of their genes in the Bacillus/Clostridium group, suggest a functional link between them. en. *J. Mol. Microbiol. Biotechnol.* **4,** 503–513 (Sept. 2002).

78. Bernard, R., Guiseppi, A., Chippaux, M., Foglino, M. & Denizot, F. Resistance to bacitracin in Bacillus subtilis: unexpected requirement of the BceAB ABC transporter in the

control of expression of its own structural genes. en. *J. Bacteriol.* **189,** 8636–8642 (Dec. 2007).

79. Kashiwagi, K., Miyamoto, S., Nukui, E., Kobayashi, H. & Igarashi, K. Functions of potA and potD proteins in spermidine-preferential uptake system in Escherichia coli. en. *J. Biol. Chem.* **268,** 19358–19363 (Sept. 1993).

80. Emerson, J. E., Stabler, R. A., Wren, B. W. & Fairweather, N. F. Microarray analysis of the transcriptional responses of Clostridium difficile to environmental and antibiotic stress. en. *J. Med. Microbiol.* **57,** 757–764 (June 2008).

81. Cairns, M. D., Preston, M. D., Lawley, T. D., Clark, T. G., Stabler, R. A. & Wren, B. W. Genomic Epidemiology of a Protracted Hospital Outbreak Caused by a Toxin A-Negative Clostridium difficile Sublineage PCR Ribotype 017 Strain in London, England. en. *J. Clin. Microbiol.* **53,** 3141–3147 (Oct. 2015).

82. Cairns, M. D., Preston, M. D., Hall, C. L., Gerding, D. N., Hawkey, P. M., Kato, H., Kim, H., Kuijper, E. J., Lawley, T. D., Pituch, H., Reid, S., Kullin, B., Riley, T. V., Solomon, K., Tsai, P. J., Weese, J. S., Stabler, R. A. & Wren, B. W. Comparative Genome Analysis and Global Phylogeny of the Toxin Variant Clostridium difficile PCR Ribotype 017 Reveals the Evolution of Two Independent Sublineages. en. *J. Clin. Microbiol.* **55,** 865–876 (Mar. 2017).

83. Stabler, R. A., Dawson, L. F., Valiente, E., Cairns, M. D., Martin, M. J., Donahue, E. H., Riley, T. V., Songer, J. G., Kuijper, E. J., Dingle, K. E. & Wren, B. W. Macro and micro diversity of Clostridium difficile isolates from diverse sources and geographical locations. en. *PLoS One* **7,** e31559 (Mar. 2012).

84. Mayer, M. J., Garefalaki, V., Spoerl, R., Narbad, A. & Meijers, R. Structure-based modification of a Clostridium difficile-targeting endolysin affects activity and host range. en. *J. Bacteriol.* **193,** 5477–5486 (Oct. 2011).

85. Monot, M., Eckert, C., Lemire, A., Hamiot, A., Dubois, T., Tessier, C., Dumoulard, B., Hamel, B., Petit, A., Lalande, V., Ma, L., Bouchier, C., Barbut, F. & Dupuy, B. Clostridium difficile: New Insights into the Evolution of the Pathogenicity Locus. en. *Sci. Rep.* **5,** 15023 (Oct. 2015).

86. Kociolek, L. K., Gerding, D. N., Hecht, D. W. & Ozer, E. A. Comparative genomics analysis of Clostridium difficile epidemic strain DH/NAP11/106. en. *Microbes Infect.* **20,** 245–253 (Apr. 2018).

87. Steiner, B. D., Eberly, A. R., Hurst, M. N., Zhang, E. W., Green, H. D., Behr, S., Jung, K. & Hadjifrangiskou, M. Evidence of Cross-Regulation in Two Closely Related Pyruvate-Sensing Systems in Uropathogenic Escherichia coli. en. *J. Membr. Biol.* **251,** 65–74 (Feb. 2018).

88. Behr, S., Fried, L. & Jung, K. Identification of a novel nutrient-sensing histidine kinase/response regulator network in Escherichia coli. en. *J. Bacteriol.* **196,** 2023–2029 (June 2014).

89. Imsande, J. Genetic regulation of penicillinase synthesis in Gram-positive bacteria. en. *Microbiol. Rev.* **42,** 67–83 (Mar. 1978).

90. Fagan, R. P., Janoir, C., Collignon, A., Mastrantonio, P., Poxton, I. R. & Fairweather, N. F. A proposed nomenclature for cell wall proteins of Clostridium difficile. en. *J. Med. Microbiol.* **60,** 1225–1228 (Aug. 2011).

91. Usenik, A., Renko, M., Mihelič, M., Lindič, N., Borišek, J., Perdih, A., Pretnar, G., Müller, U. & Turk, D. The CWB2 Cell Wall-Anchoring Module Is Revealed by the Crystal Structures of the Clostridium difficile Cell Wall Proteins Cwp8 and Cwp6. en. *Structure* **25,** 514–521 (Mar. 2017).

92. Saier Jr, M. H., Goldman, S. R., Maile, R. R., Moreno, M. S., Weyler, W., Yang, N. & Paulsen, I. T. Transport capabilities encoded within the Bacillus subtilis genome. en. *J. Mol. Microbiol. Biotechnol.* **4,** 37–67 (Jan. 2002).

93. Niehaus, T. D., Folz, J., McCarty, D. R., Cooper, A. J. L., Moraga Amador, D., Fiehn, O. & Hanson, A. D. Identification of a metabolic disposal route for the oncometabolite S-(2-succino)cysteine in Bacillus subtilis. en. *J. Biol. Chem.* **293,** 8255–8263 (May 2018).

94. Yoshida, K.-I., Fujimura, M., Yanai, N. & Fujita, Y. Cloning and Sequencing of a 23-kb Region of the Bacillus subtilis Genome between the iol and hut Operons. *DNA Res.* **2,** 295–301 (Jan. 1995).

95. Alderson, N. L., Wang, Y., Blatnik, M., Frizzell, N., Walla, M. D., Lyons, T. J., Alt, N., Carson, J. A., Nagai, R., Thorpe, S. R. & Baynes, J. W. S-(2-Succinyl)cysteine: a novel chemical modification of tissue proteins by a Krebs cycle intermediate. en. *Arch. Biochem. Biophys.* **450,** 1–8 (June 2006).

96. Yang, M., Soga, T. & Pollard, P. J. Oncometabolites: linking altered metabolism with cancer. en. *J. Clin. Invest.* **123,** 3652–3658 (Sept. 2013).

97. Frizzell, N., Thomas, S. A., Carson, J. A. & Baynes, J. W. Mitochondrial stress causes increased succination of proteins in adipocytes in response to glucotoxicity. en. *Biochem. J* **445,** 247–254 (July 2012).

98. Thomas, S. A., Storey, K. B., Baynes, J. W. & Frizzell, N. Tissue distribution of S-(2-succino)cysteine (2SC), a biomarker of mitochondrial stress in obesity and diabetes. en. *Obesity* **20,** 263–269 (Feb. 2012).

99. Saujet, L., Pereira, F. C., Serrano, M., Soutourina, O., Monot, M., Shelyakin, P. V., Gelfand, M. S., Dupuy, B., Henriques, A. O. & Martin-Verstraete, I. Genome-wide analysis of cell type-specific gene transcription during spore formation in Clostridium difficile. en. *PLoS Genet.* **9,** e1003756 (Oct. 2013).

100. Andersen, C., Rak, B. & Benz, R. The gene bglH present in the bgl operon of Escherichia coli, responsible for uptake and fermentation of beta-glucosides encodes for a carbohydrate-specific outer membrane porin. en. *Mol. Microbiol.* **31,** 499–510 (Jan. 1999).

101. Buffie, C. G., Bucci, V., Stein, R. R., McKenney, P. T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A., Littmann, E., van den Brink, M. R. M., Jenq, R. R., Taur, Y., Sander, C., Cross, J. R., Toussaint, N. C., Xavier, J. B. & Pamer, E. G. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. en. *Nature* **517,** 205–208 (Jan. 2015).

102. Jenior, M. L., Leslie, J. L., Young, V. B. & Schloss, P. D. Clostridium difficile Colonizes Alternative Nutrient Niches during Infection across Distinct Murine Gut Microbiomes. en. *mSystems* **2** (July 2017).

103. Hryckowian, A. J., Van Treuren, W., Smits, S. A., Davis, N. M., Gardner, J. O., Bouley, D. M. & Sonnenburg, J. L. Microbiota-accessible carbohydrates suppress Clostridium difficile infection in a murine model. en. *Nat Microbiol* **3,** 662–669 (June 2018).

104. Guzmán, G. I., Utrilla, J., Nurk, S., Brunk, E., Monk, J. M., Ebrahim, A., Palsson, B. O. & Feist, A. M. Model-driven discovery of underground metabolic functions in Escherichia coli. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 929–934 (Jan. 2015).

105. Ghatak, S., King, Z. A., Sastry, A. & Palsson, B. O. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. en. *Nucleic Acids Res.* **47,** 2446–2454 (Mar. 2019).

106. Poudel, S., Tsunemoto, H., Seif, Y., Sastry, A., Szubin, R., Xu, S., Machado, H., Olson, C., 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 responses* en. Mar. 2020.

107. 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).

108. Fletcher, J. R., Erwin, S., Lanzas, C. & Theriot, C. M. Shifts in the Gut Metabolome and Clostridium difficile Transcriptome throughout Colonization and Infection in a Mouse Model. en. *mSphere* **3** (Mar. 2018).

109. Midani, F. S., Collins, J. & Britton, R. A. *AMiGA: software for automated Analysis of Microbial Growth Assays* en. Nov. 2020.

110. Karasawa, T., Ikoma, S., Yamakawa, K. & Nakamura, S. A defined growth medium for Clostridium difficile. en. *Microbiology* **141 (Pt 2),** 371–375 (Feb. 1995).

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

112. 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).

113. 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).

114. Bergsma, W. A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* **42,** 323–328 (2013).

115. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. en. *Bioinformatics* **35,** 4162–4164 (Oct. 2019).

116. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. en. *Nucleic Acids Res.* **47,** W256–W259 (July 2019).

117. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. en. *Nucleic Acids Res.* **44,** W242–5 (July 2016).

118. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. en. *Bioinformatics* **31,** 3718–3720 (Nov. 2015).

119. Harris, S. R. *SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology* en. Oct. 2018.

Chapter 8

Conclusions

High throughput generation of omics datasets at unprecedented rates has resulted in an exponential increase in data availability, presenting unique big-data-to-knowledge challenges. The exponential growth of genome sequencing projects means that in one year the additional data generated will match the entirety of data that currently exists. Through the development of scalable and interpretable data modeling approaches it will be possible to convert these challenges into opportunities for greater understanding of biological systems complexity than previously possible. These data-driven outcomes will have far-reaching implications across fields from healthcare to biosustainability. Broadly, in this dissertation we study the genotype-phenotype that lies at the core of biology through developing and applying pan-genome analytics tools and models to study the diversity of microbial pathogens. In the introduction, we detailed the pangenome concept and the fruitfulness of a comparative systems biology approach enabled by dataset scale. We also illustrated the importance of genome scale reconstruction and in turn genome scale models of metabolism that allows for explorations of the phenotypic potential of a species when paired with a pangenome perspective.

The second chapter of this dissertation, "iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of *Acinetobacter baumannii* AYE" describes an updated and standardized metabolic reconstruction for *A. baumannii* AYE. iCN718 predicts gene essentiality with 80% accuracy and as much as 89% accuracy in recapitulating high-throughput growth screens. We also analyzed the conservation of metabolic functions in the species by constructing the core-genome.

The third chapter of this dissertation, "Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase-Producing Multidrug-Resistant *Klebsiella pneumoniae* Clinical Isolates" describes the use of genome scale metabolic models to generate *in silico* growth capabilities of strains with various antimicrobial resistance profiles. Within this study we demonstrated that the pan-resistome for *K. Pneumoniae* clusters according to established sequence types based only on resistance encoding mechanisms. We also used the *in silico* growth capabilities to construct classification schema for resistance profiles and found alternate nitrogen source utilization to be the best discriminator for a number of antibiotics.

The fourth chapter of this dissertation, "Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence" details high-throughput screening of three laboratory stock isolates and potential effects of intralaboratory evolution of stains. We generated and updated and improved genome scale reconstruction for *C. difficile* 630 and used this to contextualize the mutations observed across stock cultures with regard to metabolic capacity. Motivated by the divergence identified, we analyzed the alleleome of the species with 415 available strains to identify the areas of the metabolic network prone to evolution via sequence diversity.

The fifth chapter of this dissertation, "A workflow for generating multi-strain genome-

scale metabolic models of prokaryotes" details in step-by-step fashion the computational procedure to generate multi-strain models. This protocol was a critical technique utilized in the studies of the three preceding chapters and by delineating each step we have made creating multi-strain GEMs accessible to the research community. This modeling tool has the capability to scale with increasing genome sequences and offers a way to study pan-metabolic capabilities across a species and provide insight into its range of lifestyle.

The sixth chapter of this dissertation, "BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree" describes an update to the BiGG Models repository. Within the update 31 new models were added that have a demonstrated increase in the portion of the phylogenetic tree covered by the hosted models. Functionality to host multi-strain models was introduced as well as benchmarking all model content with newly available curated test metrics.

The seventh chapter of this dissertation, "Systems biology approach to functionally assess the *Clostridioides difficile* pan-genome reveals genetic diversity with discriminatory power" details a comprehensive assessment of the *C. difficile* pan-genome including the sequencing and carbon-source utilization profiling of 35 clinical isolates. We describe the development of a novel WGS-based strain type method based on the accessory genome profiles of the strains. Through this method we identified and discussed numerous cases of gene clusters that drive the separation of endemic ribotype lineages adding to the overall understanding of *C. difficile*. This technique can be applied in future studies and demonstrates the ability of data-driven methods to capture the key aspects of evolutionary trajectories within bacterial species.

Modern biotechnology has reached an inflection point where biological science is increasingly conducted at scale. Technological improvements have enabled the study of systems through

multi-omics data generation that offers comprehensive views of biology. In this dissertation we have demonstrated that pan-genome analytics offers a data-driven approach towards investigating strain properties including lifestyle, virulence, and antibiotic resistance and converting data to knowledge.

Appendix A

iCN718, an Updated and Improved Genome-Scale Metabolic Network Reconstruction of *Acinetobacter baumannii* AYE

A.1 Supplementary Figures

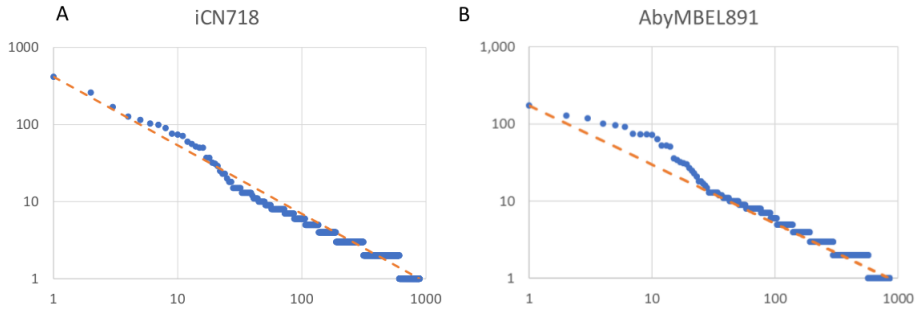

Figure A.1: for (A) iCN718 and (B) AbyMBEL891. The dashed orange line serves as a reference and points above the line indicate strong connectivity.

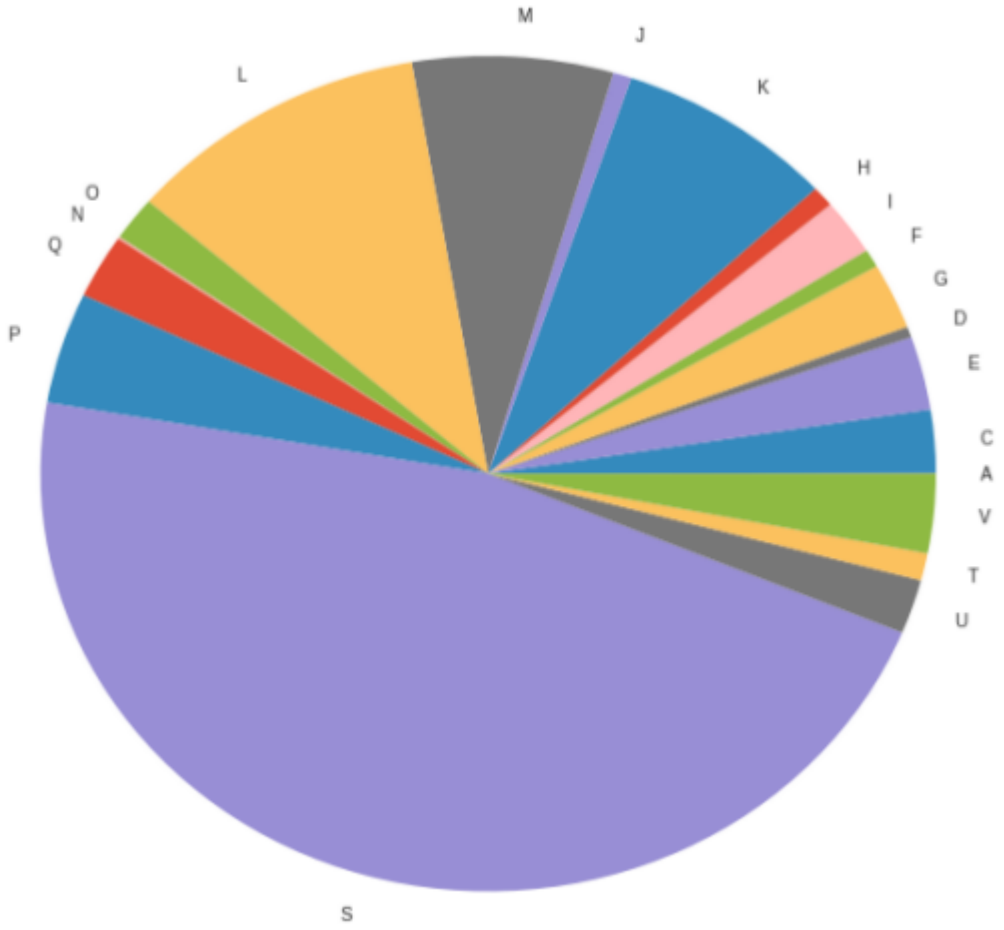

Figure A.2: of the 75 strains. Note that this only represents those genes that could be COG classified which is only half of the entire set designated in the pan-genome.

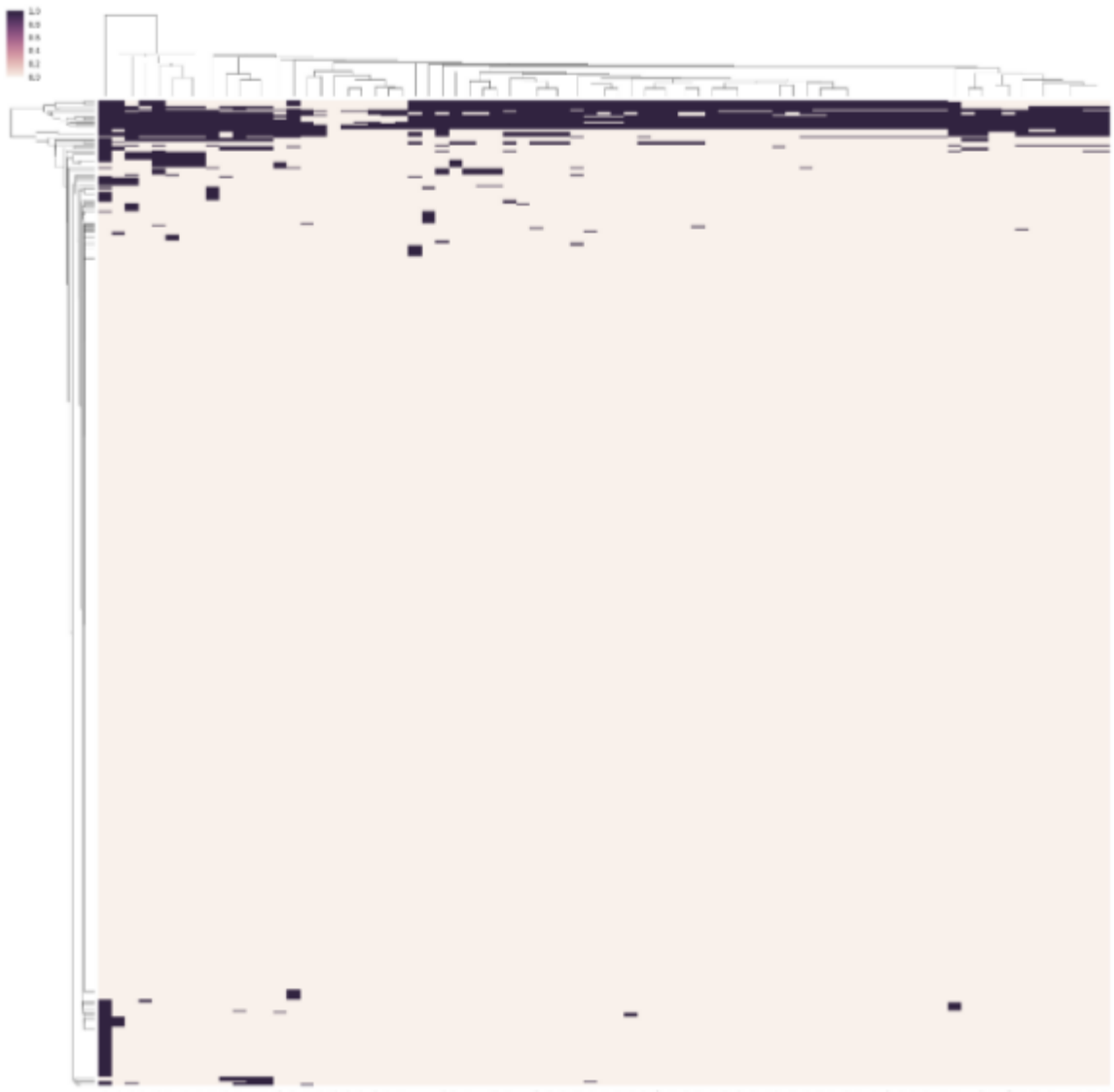

Figure A.3: Full clustermap of presence or absence for all genes in strain specific models of each of the 75 strains

A.2 Supplementary Tables

Table A.1: Comparison between carbon source Biolog Phenotypic Array data and in silico outcomes.

Compound Common Name	BiGG ID	Biolog Growth	Model Growth	Agreement
Succinate	succ_e	1	1	TP
L-Aspartate	asp__L_e	1	1	TP
D-Alanine	ala__D_e	1	1	TP
D-Xylose	xyl__D_e	1	1	TP
L-Lactate	lac__L_e	1	1	TP
L-Malate	mal__L_e	1	1	TP
Acetate	ac_e	1	1	TP
D-Glucose	glc__D_e	1	1	TP
L-Asparagine	asn__L_e	1	1	TP
2-Oxobutanoate	2obut_e	1	1	TP
Citrate	cit_e	1	1	TP
Fumarate	fum_e	1	1	TP
Propionate	ppa_e	1	1	TP
Isocitrate	icit_e	1	1	TP
L-Threonine	thr__L_e	1	1	TP
L-Alanine	ala__L_e	1	1	TP
D-Malate	mal__D_e	1	1	TP
L-Malate	mal__L_e	1	1	TP
Pyruvate	pyr_e	1	1	TP
D-Serine	ser__D_e	0	1	FP
D-Fructose	fru_e	0	1	FP
Glycolate	glyclt_e	0	1	FP
L-Serine	ser__L_e	0	1	FP
L-Arabinose	arab__L_e	1	0	FN
L-Glutamate	glu__L_e	1	0	FN
2-Oxoglutarate	akg_e	1	0	FN
L-Glutamine	gln__L_e	1	0	FN
N-Acetyl-D-glucosamine	acgam_e	0	0	TN
D-Glucarate	glcr_e	0	0	TN
D-Galactose	gal_e	0	0	TN
L-Proline	pro__L_e	0	0	TN
Trehalose	tre_e	0	0	TN
D-Mannose	man_e	0	0	TN
Galactitol	galt_e	0	0	TN
D-Sorbitol	sbt__D_e	0	0	TN

Continued on next page

Table A.1: Comparison between carbon source Biolog Phenotypic Array data and in silico outcomes, continued

Compound Common Name	BiGG ID	Biolog Growth	Model Growth	Agreement
Glycerol	glyc_e	0	0	TN
L-Fucose	fuc__L_e	0	0	TN
D-Gluconate	glcn_e	0	0	TN
Glycerol 3-phosphate	glyc3p_e	0	0	TN
Formate	for_e	0	0	TN
D-Mannitol	mnl_e	0	0	TN
D-Glucose 6-phosphate	g6p_e	0	0	TN
D-Ribose	rib__D_e	0	0	TN
L-Rhamnose	rmn_e	0	0	TN
Maltose C12H22O11	malt_e	0	0	TN
Melibiose C12H22O11	melib_e	0	0	TN
Thymidine C10H14N2O5	thymd_e	0	0	TN
D-Aspartate	asp__D_e	0	0	TN
D-Glucosamine	gam_e	0	0	TN
13-Propanediol	13ppd_e	0	0	TN
Sucrose C12H22O11	sucr_e	0	0	TN
Uridine	uri_e	0	0	TN
L-tartrate	tartr__L_e	0	0	TN
D-Glucose 1-phosphate	g1p_e	0	0	TN
D-Fructose 6-phosphate	f6p_e	0	0	TN
Beta-Methylglucoside	mbdg_e	0	0	TN
Ribitol	rbt_e	0	0	TN
Maltotriose	malttr_e	0	0	TN
Deoxyadenosine	dad_2_e	0	0	TN
Adenosine	adn_e	0	0	TN
Myo-Inositol	inost_e	0	0	TN
D-Galactarate	galct__D_e	0	0	TN
Glyoxylate	glx_e	0	0	TN
Inosine	ins_e	0	0	TN
Gly glu L	gly_glu__L_e	0	0	TN
Acetoacetate	acac_e	0	0	TN
N-Acetyl-D-mannosamine	acmana_e	0	0	TN
4 Hydroxyphenylacetic acid	4hoxpac_e	0	0	TN
3 Hydroxyphenylacetic acid	3hoxpac_e	0	0	TN
Tyramine	tym_e	0	0	TN
D-Allose	all__D_e	0	0	TN
L-Lyxose	lyx__L_e	0	0	TN

Continued on next page

Table A.1: Comparison between carbon source Biolog Phenotypic Array data and in silico outcomes, continued

Compound Common Name	BiGG ID	Biolog Growth	Model Growth	Agreement
D-Galacturonate	galur_e	0	0	TN
Phenethylamine	peamn_e	0	0	TN

Table A.2: Comparison between nitrogen source Biolog Phenotypic Array data and in silico outcomes.

Compound Common Name	BiGG ID	Biolog Growth	Model Growth	Agreement
Ammonium	nh4_e	1	1	TP
Nitrite	no2_e	1	1	TP
Nitrate	no3_e	1	1	TP
Urea	urea_e	1	1	TP
L-Alanine	ala__L_e	1	1	TP
L-Arginine	arg__L_e	1	1	TP
L-Asparagine	asn__L_e	1	1	TP
L-Aspartate	asp__L_e	1	1	TP
L-Glutamate	glu__L_e	1	1	TP
L-Glutamine	gln__L_e	1	1	TP
Glycine	pro__L_e	1	1	TP
L-Proline	ser__L_e	1	1	TP
L-Serine	ala__D_e	1	1	TP
L-Threonine	glu__D_e	1	1	TP
D-Alanine	orn_e	1	1	TP
D-Glutamate	etha_e	0	1	FP
D-Serine	gly_e	0	1	FP
Ornithine	ser__D_e	0	1	FP
Ethanolamine	thr__L_e	0	1	FP
L-Cysteine	cys__L_e	0	0	TN
L-Isoleucine	ile__L_e	0	0	TN
L-Leucine	leu__L_e	0	0	TN
L-Lysine	lys__L_e	0	0	TN
L-Methionine	met__L_e	0	0	TN
L-Phenylalanine	phe__L_e	0	0	TN
L-Tryptophan	trp__L_e	0	0	TN
L-Valine	val__L_e	0	0	TN
D-Aspartate	asp__D_e	0	0	TN

Continued on next page

Table A.2: Comparison between nitrogen source Biolog Phenotypic Array data and in silico outcomes, continued

Compound Common Name	BiGG ID	Biolog Growth	Model Growth	Agreement
Putrescine	ptrc_e	0	0	TN
Tyramine	tym_e	0	0	TN
D-Glucosamine	gam_e	0	0	TN
N-Acetyl-D-mannosamine	acmana_e	0	0	TN
Adenosine	adn_e	0	0	TN
Cytidine	cytd_e	0	0	TN
Cytosine	csn_e	0	0	TN
Guanosine	gsn_e	0	0	TN
Thymidine C10H14N2O5	thymd_e	0	0	TN
Uracil	ura_e	0	0	TN
Uridine	uri_e	0	0	TN
Inosine	ins_e	0	0	TN
Xanthine	xan_e	1	0	FN
L-Tyrosine	tyr__L_e	1	0	FN
L-Histidine	his__L_e	1	0	FN

Table A.3: Simmons minimal media composition in silico.

Reaction Name	Compound	Lower Bound	Upper Bound
EX_h2o_e	h2o	-10	1000
EX_o2_e	o2	-10	1000
EX_so4_e	so4	-10	1000
EX_nh4_e	nh4	-10	1000
EX_mg2_e	mg2	-10	1000
EX_na1_e	na1	-10	1000
EX_cl_e	cl	-10	1000
EX_pi_e	pi	-10	1000
EX_h_e	h	-10	1000
EX_cit_e	cit	-10	1000

Table A.4: Synthetic lethal gene pairs.

Gene 1	Gene 2
ABAYE3800	ABAYE2928
ABAYE0899	ABAYE1650
ABAYE0899	ABAYE1510
ABAYE2088	ABAYE3443
ABAYE0781	ABAYE0783
ABAYE0781	ABAYE0784
ABAYE0783	ABAYE0780
ABAYE0784	ABAYE0780
ABAYE1650	ABAYE1510
ABAYE2116	ABAYE2823
ABAYE2116	ABAYE2824
ABAYE1562	ABAYE3804
ABAYE2053	ABAYE2783
ABAYE2809	ABAYE0262
ABAYE2227	ABAYE2993
ABAYE1223	ABAYE0817
ABAYE2981	ABAYE1789
ABAYE1280	ABAYE3887
ABAYE1366	ABAYE3887
ABAYE0912	ABAYE3887
ABAYE0889	ABAYE3887
ABAYE0056	ABAYE3887
ABAYE1039	ABAYE3887
ABAYE1367	ABAYE3887
ABAYE1367	ABAYE0645
ABAYE2592	ABAYE3887
ABAYE2592	ABAYE0062
ABAYE0166	ABAYE3887
ABAYE0166	ABAYE0645
ABAYE3740	ABAYE1456
ABAYE3661	ABAYE2940
ABAYE0935	ABAYE2838
ABAYE2596	ABAYE3293
ABAYE2596	ABAYE0264
ABAYE3348	ABAYE3293
ABAYE3348	ABAYE0264
ABAYE2822	ABAYE3293
ABAYE2822	ABAYE0264
ABAYE3696	ABAYE0645

Continued on next page

Table A.4: Synthetic lethal gene pairs, continued

Gene 1	Gene 2
ABAYE1658	ABAYE1989
ABAYE1682	ABAYE1658
ABAYE1539	ABAYE1682
ABAYE0379	ABAYE0096
ABAYE0096	ABAYE2987
ABAYE0916	ABAYE2666
ABAYE0812	ABAYE3887
ABAYE0812	ABAYE0645
ABAYE2630	ABAYE3678
ABAYE0062	ABAYE1039

Table A.5: Genome IDs and strain names used for pan-genome analysis.

Genome ID	Strain Name
470.1311	Acinetobacter baumannii strain CR17
470.771	Acinetobacter baumannii AbH12O-A2
1400867.3	Acinetobacter baumannii ZW85-1
1401639.4	Acinetobacter baumannii NCGM 237
1413216.3	Acinetobacter baumannii AB07
1100841.3	Acinetobacter baumannii TYTH-1
400667.7	Acinetobacter baumannii ATCC 17978
980514.3	Acinetobacter baumannii TCDC-AB0715
696749.3	Acinetobacter baumannii 1656-2
470.774	Acinetobacter baumannii IOMTU 433
470.1822	Acinetobacter baumannii strain YU-R612
470.2928	Acinetobacter baumannii strain CMC-CR-MDR-Ab4
470.2929	Acinetobacter baumannii strain CMC-MDR-Ab59
470.3044	Acinetobacter baumannii strain AB042
470.2423	Acinetobacter baumannii strain 3027STDY5784958
470.1738	Acinetobacter baumannii
470.775	Acinetobacter baumannii A1
557600.4	Acinetobacter baumannii AB307-0294
1455315.5	Acinetobacter baumannii LAC-4
470.2917	Acinetobacter baumannii strain LAC4
509170.6	Acinetobacter baumannii SDF
509173.8	Acinetobacter baumannii AYE
470.3774	Acinetobacter baumannii strain A85

Continued on next page

Table A.5: Genome IDs and strain names used for pan-genome analysis, continued

Genome ID	Strain Name
480119.5	Acinetobacter baumannii AB0057
405416.6	Acinetobacter baumannii ACICU
497978.4	Acinetobacter baumannii MDR-ZJ06
470.2913	Acinetobacter baumannii strain XDR-BJ83
1096995.4	Acinetobacter baumannii BJAB07104
1096996.4	Acinetobacter baumannii BJAB0715
1096997.4	Acinetobacter baumannii BJAB0868
470.1345	Acinetobacter baumannii strain AB5075-UW
470.1405	Acinetobacter baumannii strain D36
470.2122	Acinetobacter baumannii strain 3207
470.2911	Acinetobacter baumannii strain AF-673
470.1864	Acinetobacter baumannii strain XH860
470.1865	Acinetobacter baumannii strain XH859
470.2912	Acinetobacter baumannii strain AF-401
470.2026	Acinetobacter baumannii strain Ab421_GEIH-2010
470.1295	Acinetobacter baumannii strain AB30
470.1294	Acinetobacter baumannii strain AB31
470.1866	Acinetobacter baumannii strain XH857
470.1867	Acinetobacter baumannii strain XH856
470.1869	Acinetobacter baumannii strain XH858
470.1288	Acinetobacter baumannii strain AC29
470.773	Acinetobacter baumannii 6200
470.1576	Acinetobacter baumannii strain Ab04-mff
470.1763	Acinetobacter baumannii strain KBN10P02143
470.2931	Acinetobacter baumannii strain 11510
470.1737	Acinetobacter baumannii
470.3354	Acinetobacter baumannii strain WKA02
470.3353	Acinetobacter baumannii strain HWBA8
470.3351	Acinetobacter baumannii strain USA2
470.3352	Acinetobacter baumannii strain SSA6
470.3347	Acinetobacter baumannii strain JBA13
470.3348	Acinetobacter baumannii strain CBA7
470.3358	Acinetobacter baumannii strain SSMA17
470.3357	Acinetobacter baumannii strain SSA12
470.3355	Acinetobacter baumannii strain USA15
470.3356	Acinetobacter baumannii strain SAA14
470.1574	Acinetobacter baumannii strain B8300
470.1579	Acinetobacter baumannii strain B8342
470.1375	Acinetobacter baumannii strain XH386
470.1575	Acinetobacter baumannii strain ATCC 17978-mff

Continued on next page

Table A.5: Genome IDs and strain names used for pan-genome analysis, continued

Genome ID	Strain Name
470.1765	Acinetobacter baumannii
470.2908	Acinetobacter baumannii strain HRAB-85
470.3349	Acinetobacter baumannii strain 15A34
470.3106	Acinetobacter baumannii strain AB34299
470.2668	Acinetobacter baumannii strain KAB01
470.2669	Acinetobacter baumannii strain KAB02
470.2671	Acinetobacter baumannii strain KAB04
470.2672	Acinetobacter baumannii strain KAB05
470.2673	Acinetobacter baumannii strain KAB06
470.2674	Acinetobacter baumannii strain KAB07
470.2675	Acinetobacter baumannii strain KAB08
470.3362	Acinetobacter baumannii strain ab736

Appendix B

Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase-Producing Multidrug-Resistant *Klebsiella pneumoniae* Clinical Isolates.

B.1 Supplementary Figures

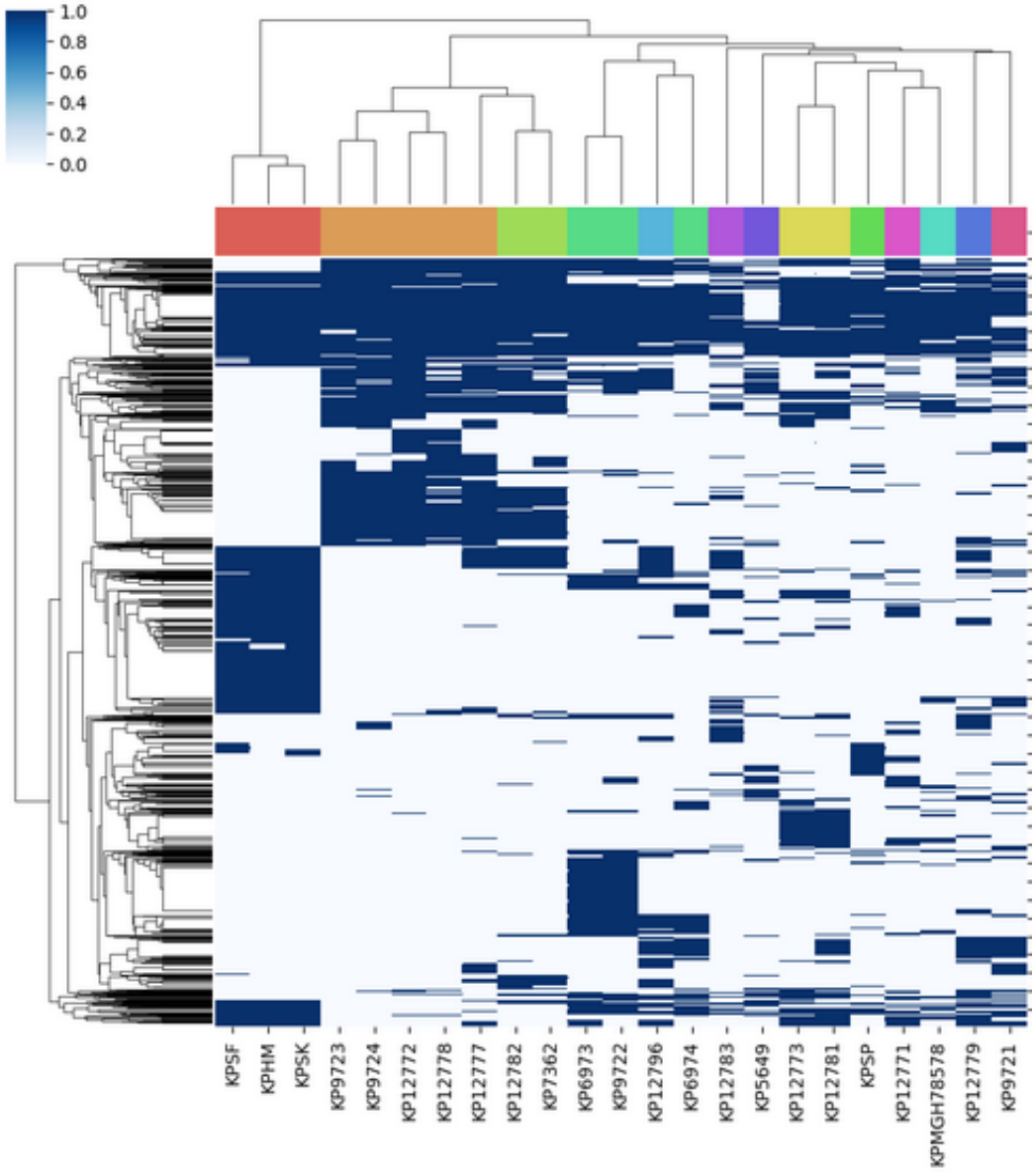

Figure B.1: Hierarchical clustering of the accessory genomes of 22 *K. pneumoniae* strains

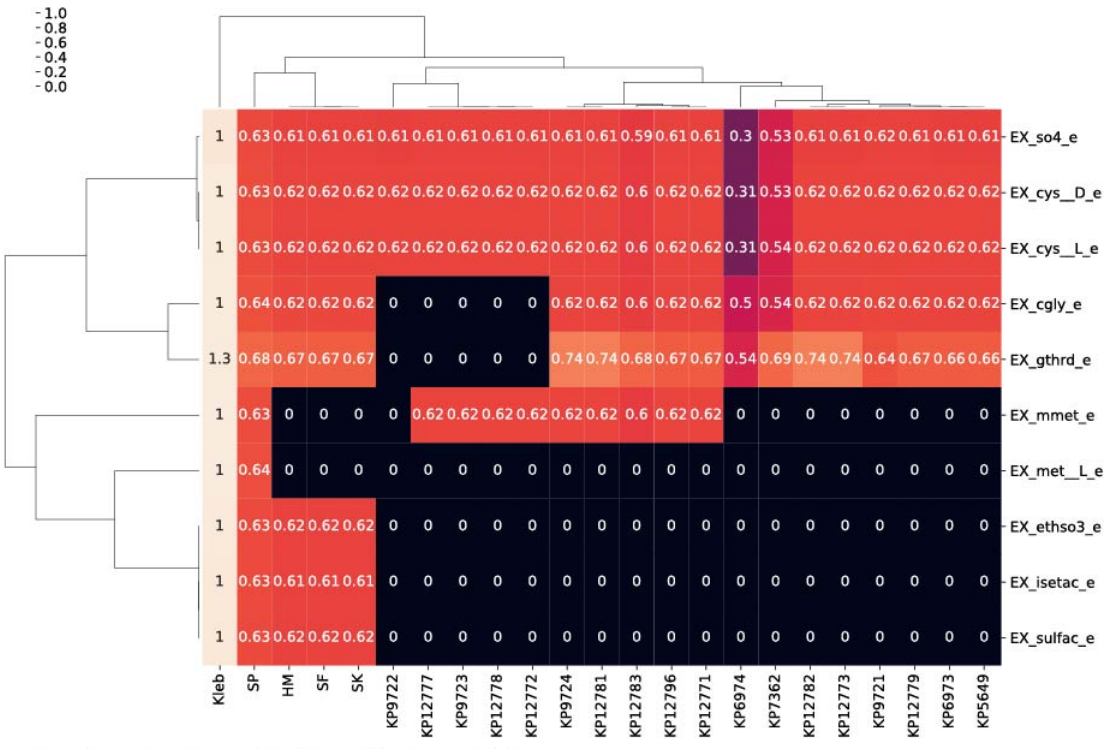

Figure B.2: Sulfur Catabolic Capabilities

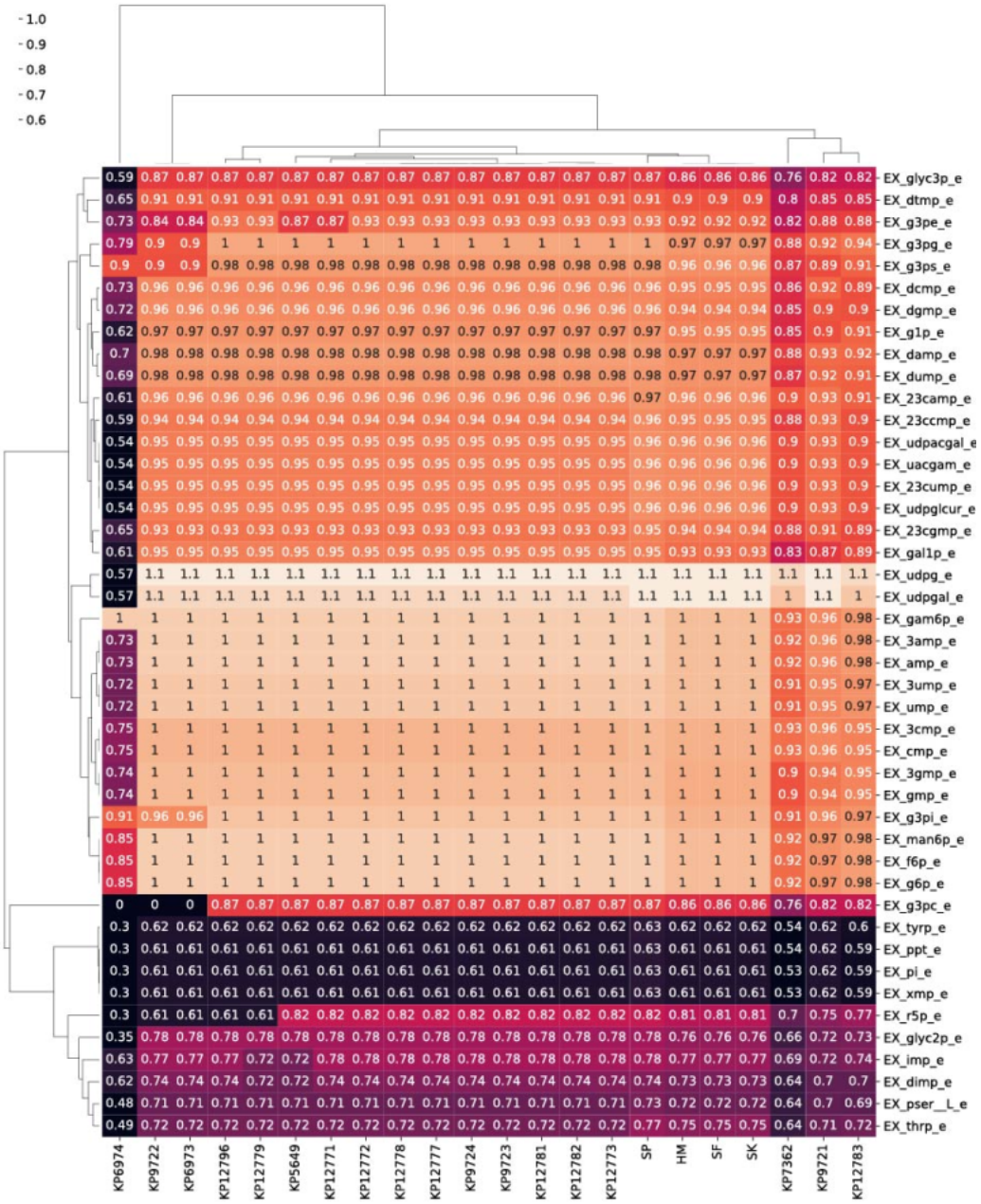

Figure B.3: Phosphorus Catabolic Capabilities

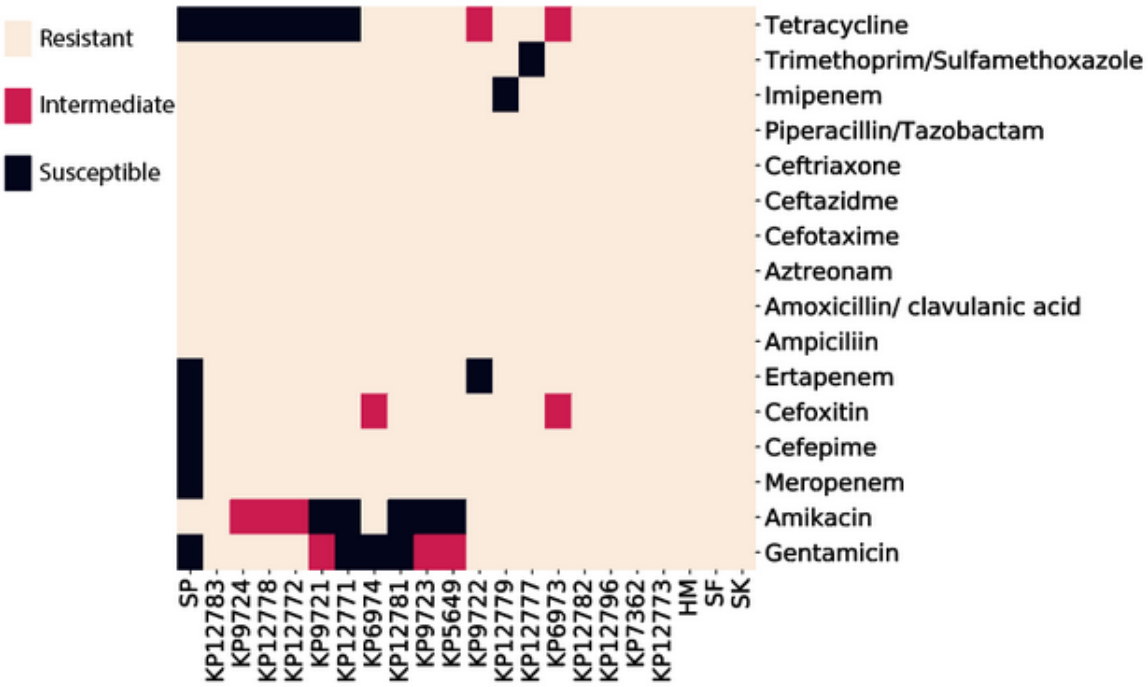

Figure B.4: Antimicrobial Resistance Profiles

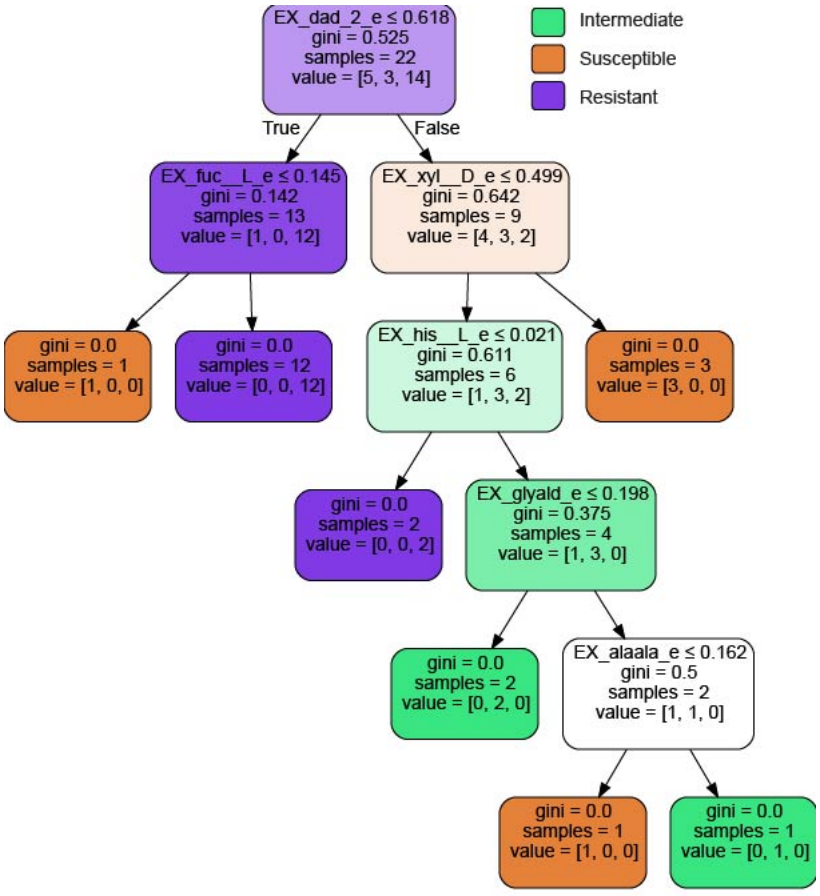

Figure B.5: Classification tree built for 22 strains on carbon source utilization for amikacin phenotypes

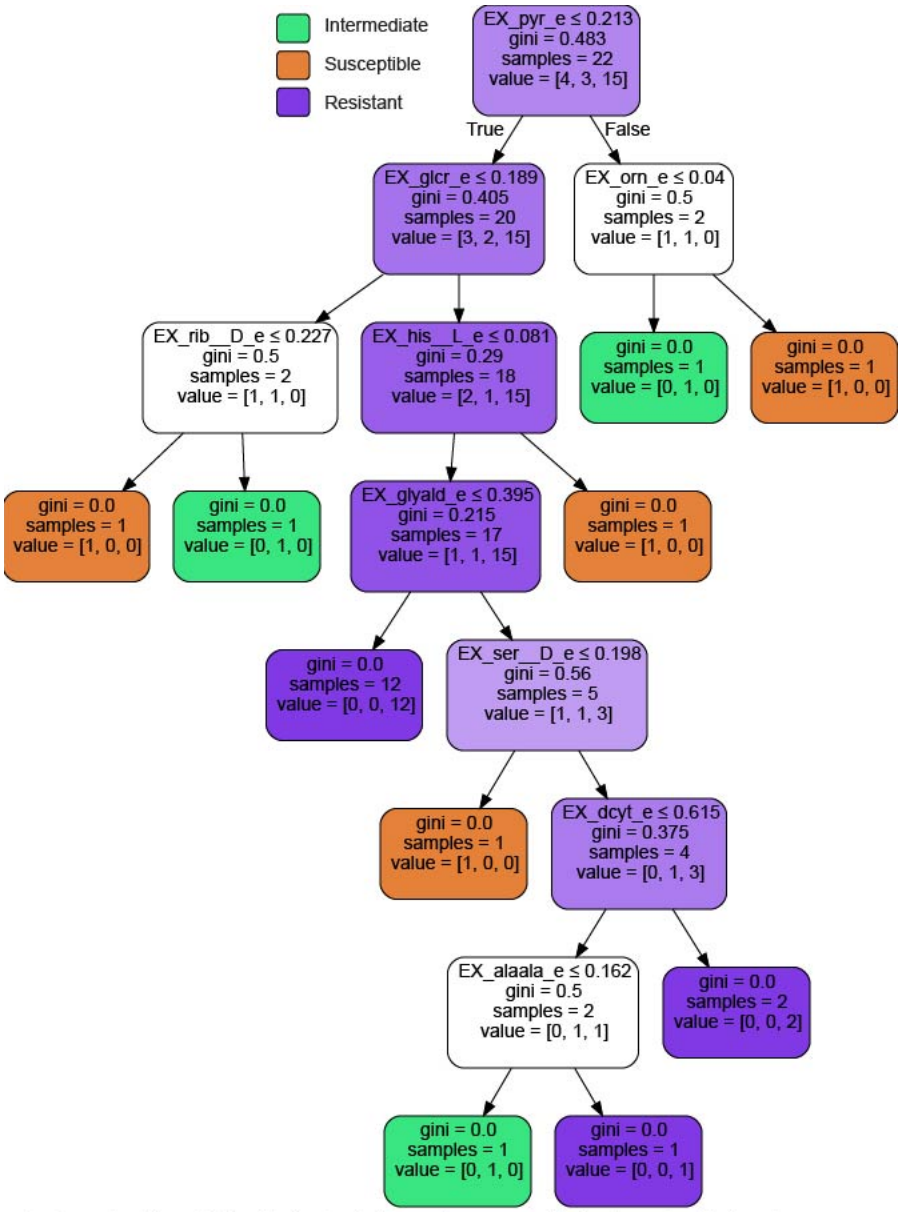

Figure B.6: Classification tree built for 22 strains on carbon source utilization for gentamicin phenotypes

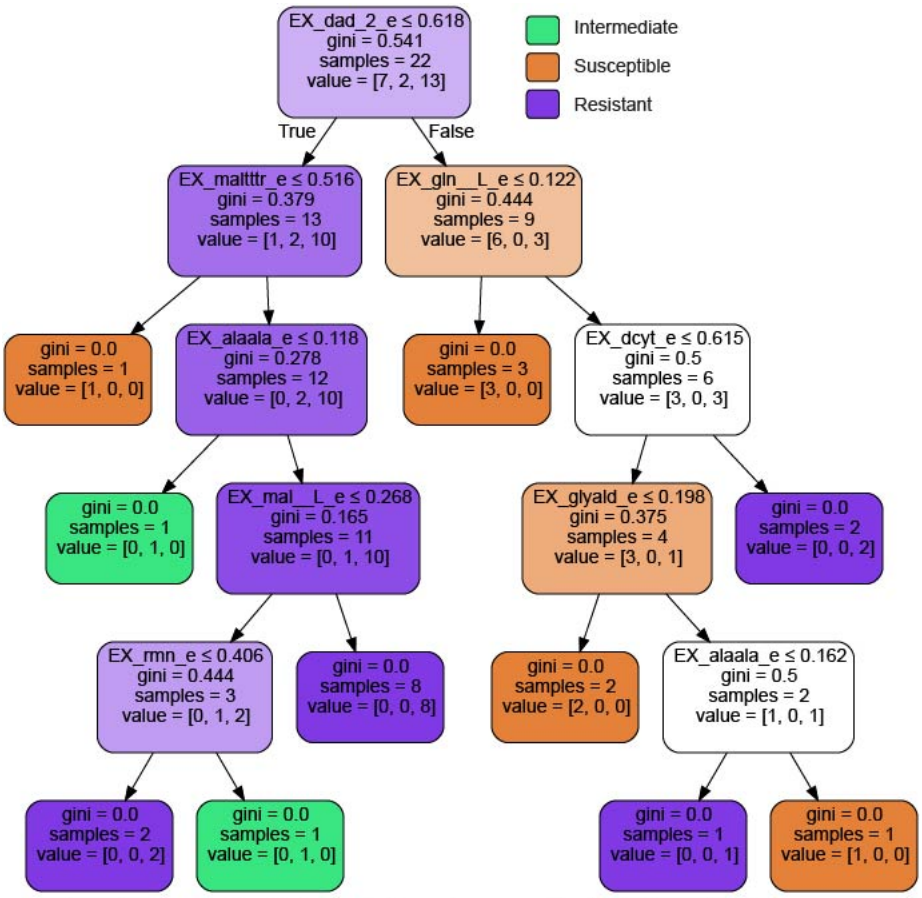

Figure B.7: Classification tree built for 22 strains on carbon source utilization for tetracycline phenotypes

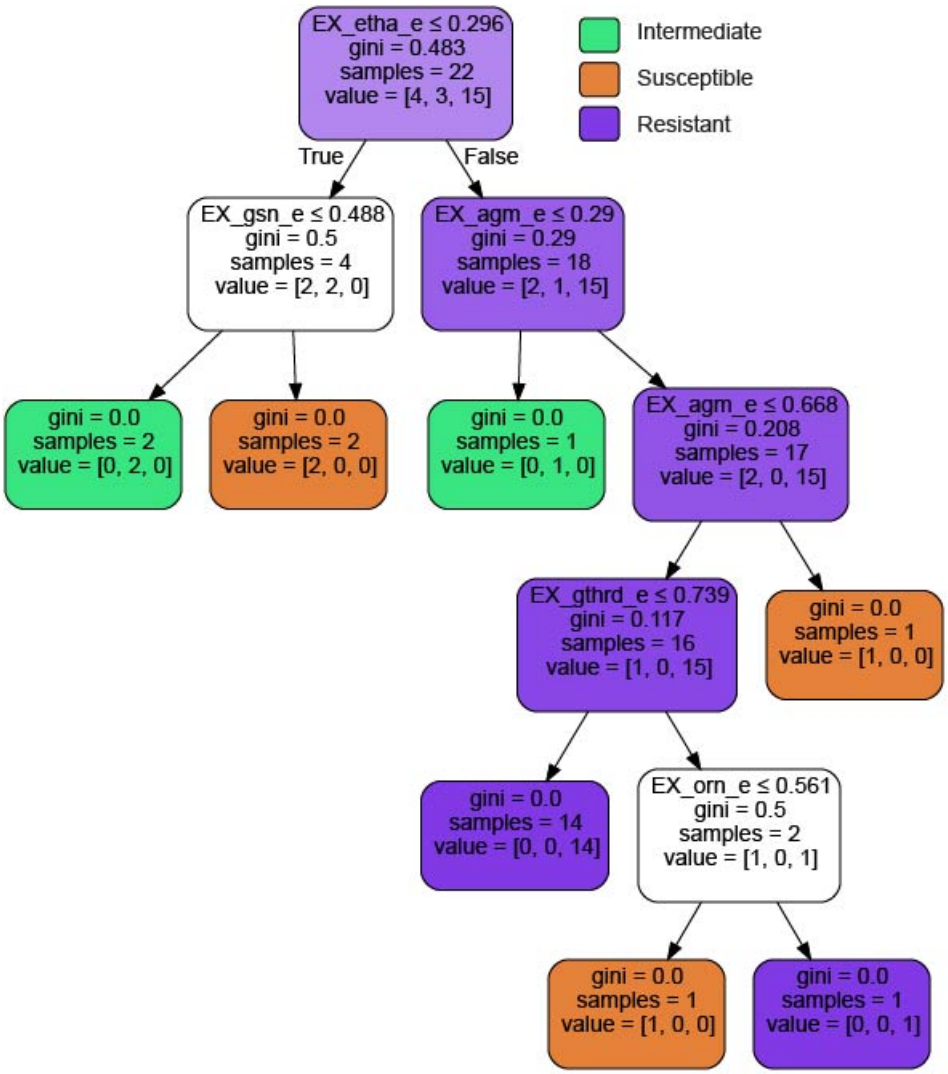

Figure B.8: Classification tree built for 22 strains on nitrogen source utilization for gentamicin phenotypes

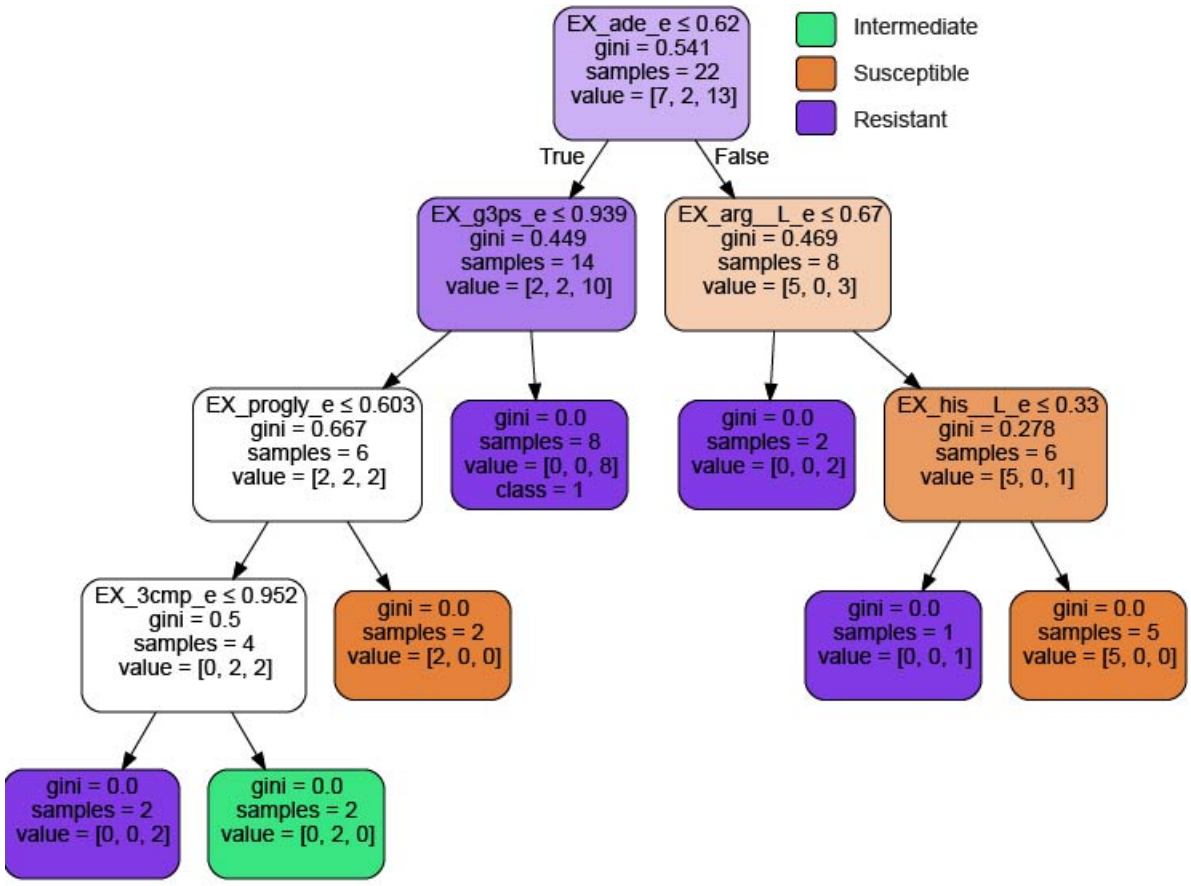

Figure B.9: Classification tree built for 22 strains on nitrogen source utilization for tetracycline phenotypes

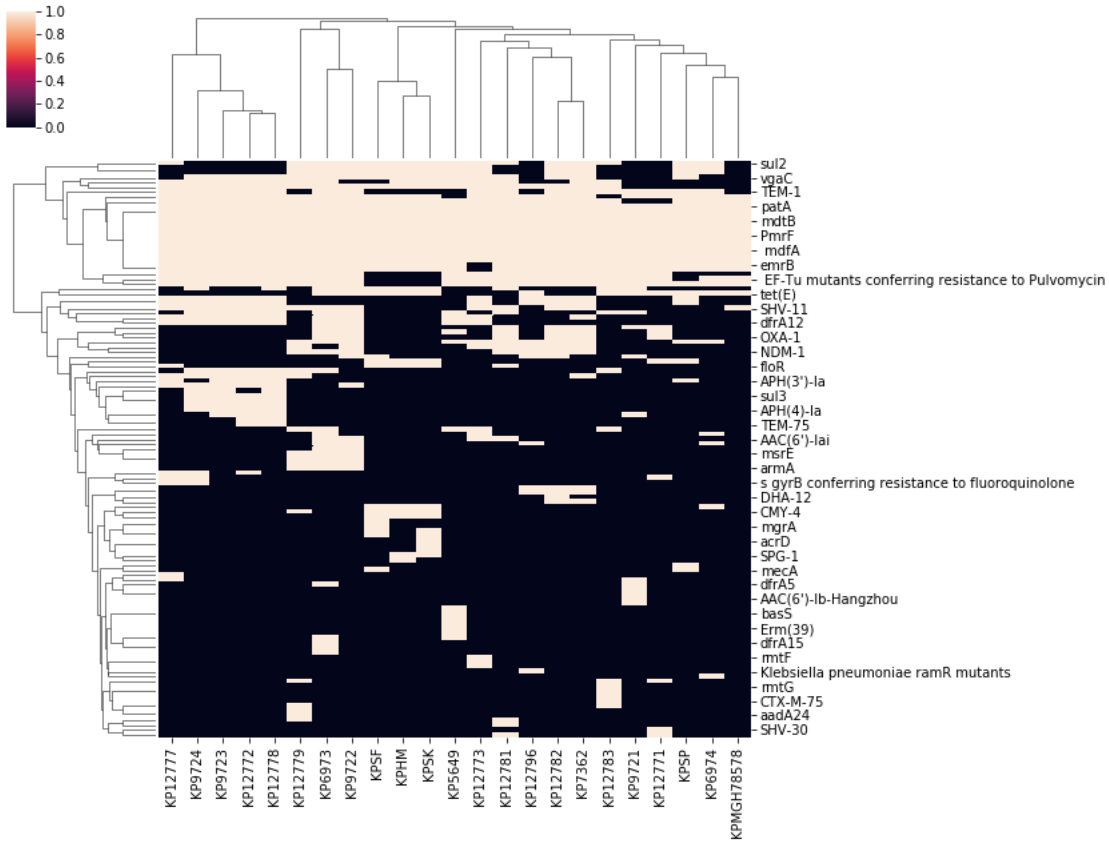

Figure B.10: Resistance Determinants for All Strains

B.2 Supplementary Tables

Table B.1: The MIC (µg/ml) of the antibiotics against the selected four isolates.

Isolate	Ceftazidime	Cefotaxime	Meropenem	Ertapenem	Colistin
HH	>256	>256	>8	>64	1
SF	>256	>256	>8	>64	4
SK	32	>256	4	>64	4
SP	>256	>256	>8	>64	2

Table B.2: The susceptibility of the isolates to these antibiotics was determined according to MIC interpretation chart, where MIC Interpretative Standard (µg/ml)

Antimicrobial agent	S	I	R
Cefotaxime	≤ 1	2	≥ 4
Ceftazidime	≤ 4	8	≥ 16
Colistin	≤ 1	2	≥ 4
Ertapenem	≤ 0.25	0.5	≥ 1
Meropenem	≤ 1	2	≥ 4

Table B.3: In Silico Media Composition

Compound	Exchange Reaction	Lower Bound
Glucose	EX_glc__D_e	-10
Calcium	EX_ca2_e	-10
Chloride	EX_cl_e	-10
Carbon Dioxide	EX_co2_e	-10
Cobalt	EX_cobalt2_e	-10
Copper	EX_cu2_e	-10
Iron	EX_fe2_e	-10
Hydrogen	EX_h_e	-10
Magnesium	EX_mg2_e	-10
Manganese	EX_mn2_e	-10
Molybdate	EX_mobd_e	-10
Sodium	EX_na1_e	-10
Ammonia	EX_nh4_e	-10
Oxygen	EX_o2_e	-10
Phosphate	EX_pi_e	-10
Sulfate	EX_so4_e	-10
Tungstate	EX_tungs_e	-10
Zinc	EX_zn2_e	-10

Appendix C

Systems biology analysis of the *Clostridioides difficile* core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence

C.1 Supplementary Text

C.1.1 Model Reconstruction Process

We began the network reconstruction by evaluating both the existing GEMs for *C. difficile* 630: iMLTC806cdf and icdf834. The first reconstruction produced by Laroque et. al in 2014 included 806 genes, 1,013 reactions, and 703 metabolites and represented the first effort at a manually curated network reconstruction for *C. difficile*. In addition to the first curated network this work included validation of the model on four types of *in silico* media, identification of essential amino acids, and evaluation in comparison to an automatically generated network. In 2017 Kashaf et al. produced icdf834 a second GEM that improved upon iMLTC806cdf. icdf834 includes 834 genes, 1227 reactions, and 807 metabolites and the major expansion of content is reflected in the inclusion of fatty acid, glycerolipid, and glycerophospholipid pathways. Overall, iMLTC806cdf and icdf834 provided a valuable starting point and iCN900 represents the next step in this lineage providing increases in both network quality and content. To expedite the process of improving and adding to the network we first translated the previous efforts to standardized BiGG format for reaction and metabolite identifiers. By putting the model into a standardized notation the tractability of the network has been greatly improved and now the iCN900 is a part of a large repository of GEMs in BiGG notation that includes a diverse phylogeny of organisms. The slight drop in reaction number from icdf834 to iCN900 is a result of changing the duplicate secretion and exchange reaction set up in icdf834 to a more conventional set of single exchange reactions. We standardized the previous networks to BiGG 19,20 reaction and metabolite identifiers to increase the usability of the reconstruction. We also subjected the previous curated network to rigorous validation and removed the presence of many erroneous energy generating cycles. Building upon

the now robust version of the metabolic network derived from the foundation of iMLTC806cdf and icdf834 we added a significant amount of new content to the reconstruction. As a means of further quality assurance of iCN900 we ran the reconstruction through the MEMOTE test suite for consistency. iCN900 scores 100% on the metrics of charge balance, metabolite connectivity, and checks for unbounded flux in default medium. iCN900 scores 87.1% for mass balance and 17% for stoichiometric consistency. We hypothesized that the lack of mass information for some components of the lipid metabolism results in this issue. These reactions have not been fully characterized and the ambiguous stoichiometry likely results in this one particularly low score. To test this we created a version of the model excluding these such reactions and saw the expected increase in the score for stoichiometric consistency. Thus this is less an issue of the reconstruction and represents a current knowledge gap on the composition of these metabolites and reaction stoichiometries.

iCN900 contains an additional 66 genes, 46 reactions, and 70 metabolites versus the content present in icdf834. These additions were made through a variety of techniques including use of the annotation tool DETECT v2, BLAST with the most closely related reconstructions, and curation of pathways based on false negative model predictions against experimental data. DETECT v2 is an enzyme annotation tool that assigns potential enzyme commission number to protein sequence. We ran DETECT v2 on the reference genome for *C. difficile* 630 and extensively looked through the results cross-referencing with the genome annotation to find a number of new genes with predicted metabolic function that could be added to the model. This proved to be a valuable method for identifying candidate new reactions and corresponding gene product rules (GPRs), but necessitated rigorous examination of the automatic results to ensure accuracy. The second means for adding content to the reconstruction was by utilizing BLAST to

identify homologous genes with *Bacillus subtilis* and *Clostridium ljungdahlii*. These organisms were chosen on the basis that they are the most closely related organisms for which there exists a high-quality GEM. Utilizing the homologous genes and GPRs from these reconstructions we were able to fill gaps in the previous *C. difficile* network. Encouragingly, many of the genes that were eventually added were identified independently in both the DETECT v2 and BLAST homology based workflows. Lastly, there were a number of compounds from experiments with Biolog Phenotypic Microarrays that the model originally incorrectly predicted unable to sustain growth. These false negative predictions provide opportunity for further network curation since the experimental data suggests the organism has the necessary machinery to grow on these compounds. Thus iCN900 includes reaction content that reconciles three key false negative model predictions for salicin, arbutin, and N-acetyl-galactosamine into agreement with the experimental data. The reactions included for each are SALCpts and S6PG, ARBTpts and AB6PGH, and ACGALpts and ACGAL6PI, which are all gene annotated with the exception of ACGAL6PI.

C.1.2 Model Validation

Essential gene predictions were performed as had been done previously with iMLTC806cdf and icdf834. Critically, in the evaluation of icdf834, Kashaf et al. utilize the experimental dataset of essential genes for *C. difficile* R20291 that had been generated in the intervening time between iMLTC806cdf and icdf834. The switch from comparison to *Bacillus subtilis* essential gene data was conducted originally for iMLTC806cdf was a significant improvement.

iCN900 had 90% accuracy prediction of essential genes compared to the Dembek dataset. While Kashaf et al. report an accuracy for icdf834 of 92.3% they calculate their accuracy only on the predicted model essential genes and not the full confusion matrix of homologous genes.

We reran gene essentiality predictions with icdf834 and found the overall accuracy in comparison to the R20291 experimental data to also be 90%, but for a smaller number of homologous genes since iCN900 reflects an increase in gene content. Overall the predictions made with iCN900 are based on 868 homologous genes to R20291 and evaluation of the full confusion matrix results in a Matthews correlation coefficient of .504.

Of the 190 compounds screened in the assay 114 of them directly map to metabolites within the BiGG database. Each of the profiled strains was compared to the in silico predicted growth capabilities and resulted in the following accuracies on tractable metabolites: 74.56%, 72.8%, and 67.5%. If the assumption is made that the model would predict no growth on the compounds that do not map to BiGG and are therefore not included within the model, then the accuracies increase to 81%, 80%, and 78%. Overall, iCN900 demonstrates a high prediction accuracy to the phenotypic data.

C.1.3 Further False Negative and False Positive Predictions

Leucine and methionine were also classified originally as false negatives, but we hypothesize that this is an artifact of recapitulating the proprietary Biolog media as defined *in silico* media. As both these metabolites are already within the minimal media, we looked at the relative biomass yield if the amount of either of these compounds was increased. From this analysis we found that increasing the leucine available resulted in increased biomass yield whereas increased methionine did not. As such leucine may be considered a true positive by the model and methionine remains a false negative. Additionally D-arabitol is likely a false negative that is more accurately considered a true negative prediction as the fold change in OD from the Biolog experiment is right at the threshold in our analyses for what we consider growth. This is

supported by the fact that the growth values for L-arabitol definitively show no growth. There are two remaining false negative model predictions that could not be rectified to true positives that merit further analysis: succinate and ethanolamine.

Previous studies suggest that *C. difficile* 630 can utilize succinate as a carbon source through the usage of succinate to butyrate pathway. While iCN900 includes the aforementioned pathway and corresponding supplemental pathways such as sorbitol fermentation pathway, thought to provide complimentary electron flow, the model predicts no growth when succinate is the sole carbon source in minimal media. It is worth noting that the addition of a succinate dehydrogenase using ubiquinone as a cofactor would enable growth on succinate, however there is no compelling genetic basis for this reaction and therefore it was not added.

C.2 Supplementary Figures

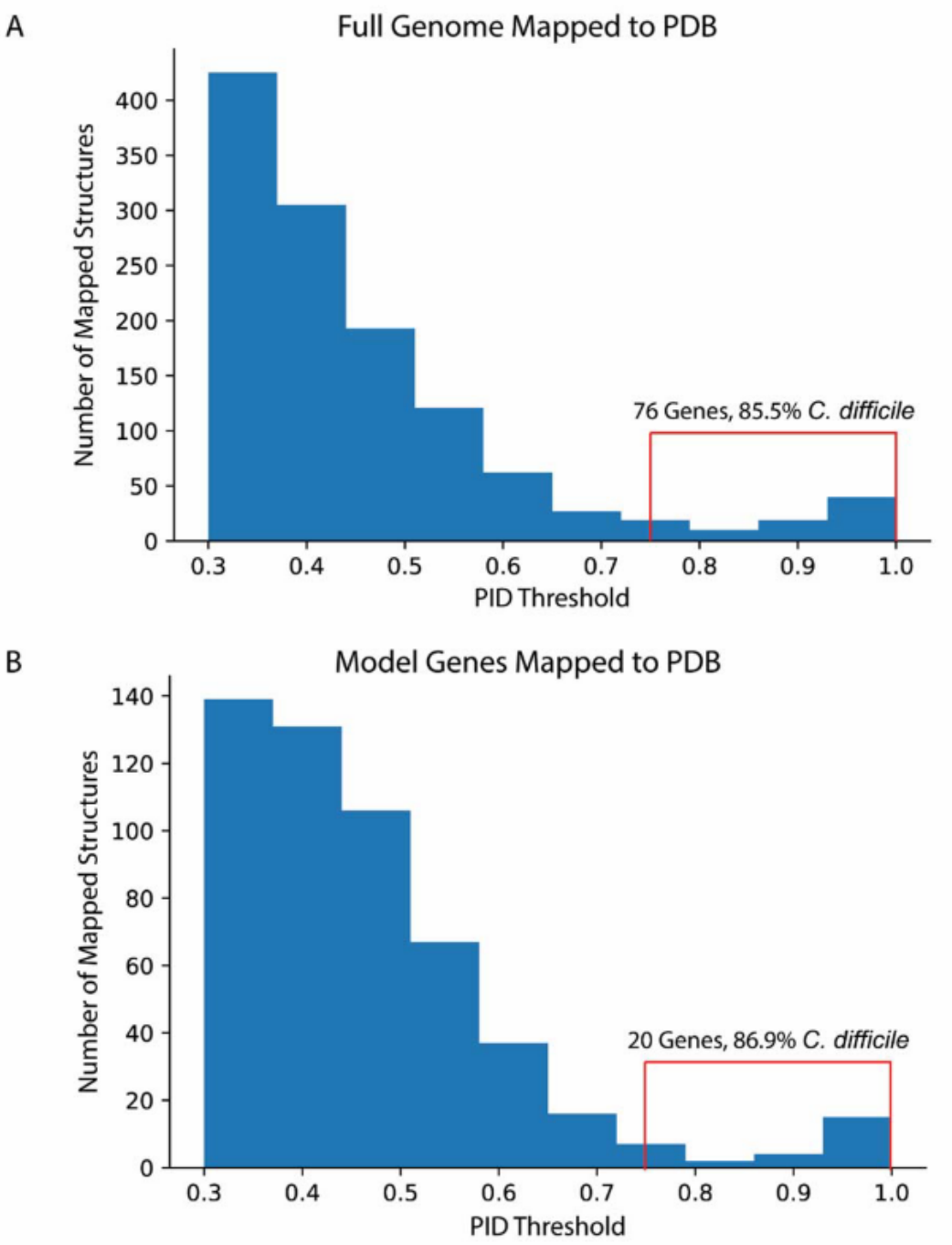

Figure C.1: Histogram detailing the amount of genes mapped to the PDB within the full reference genome and within iCN900 model genes

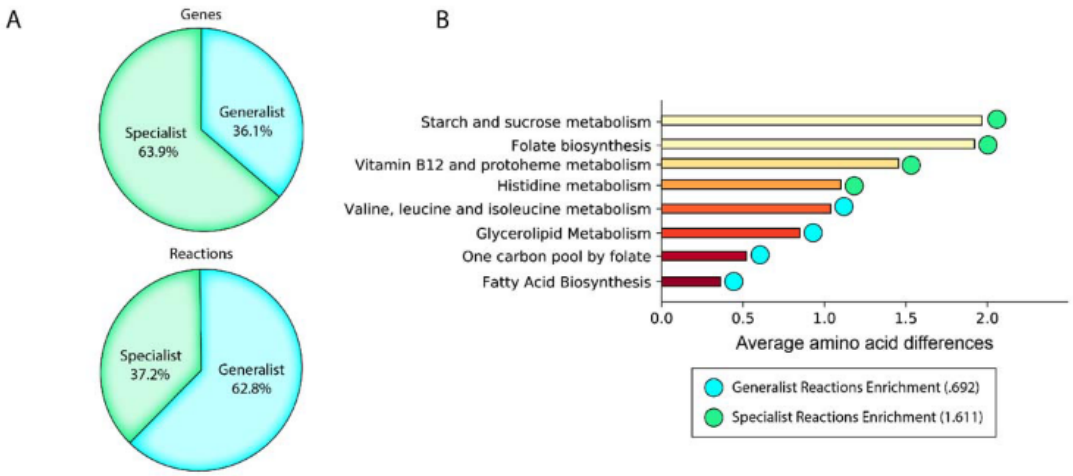

Figure C.2: A) Pie chart detailing the percentage of genes and reactions designated as either specialists or generalists. B) Each subsystem was checked for significant enrichment via hypergeometric test and the subsystems with enrichments are shown along with the corresponding average amino acid sequence variation.

Appendix D

A workflow for generating multi-strain genome-scale metabolic models of prokaryotes

D.1 Supplementary Figures

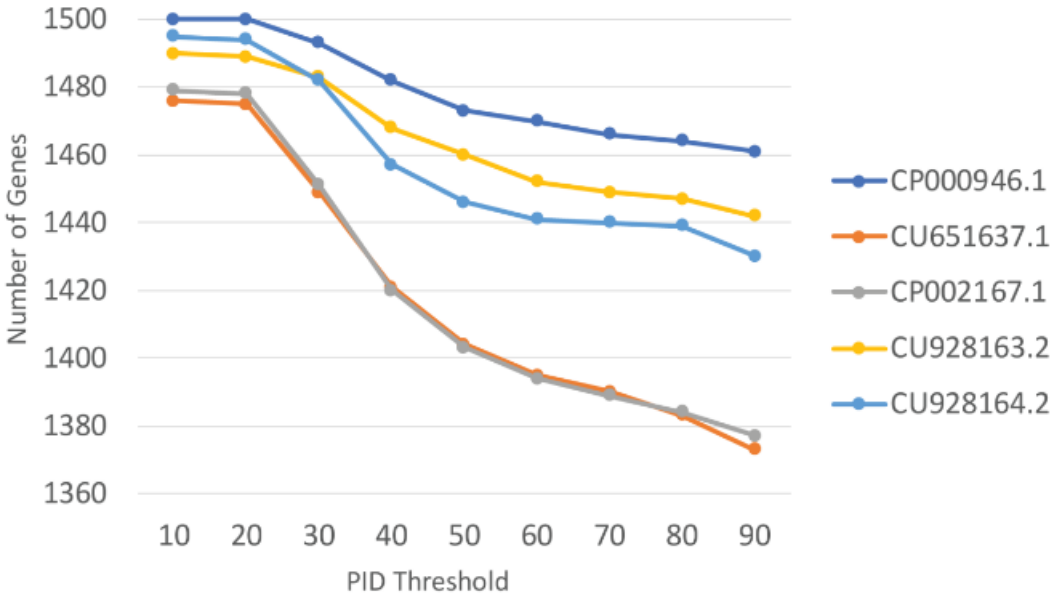

Figure D.1: The numberof genes retained in each strain-specific model is dependent on the threshold utilized for binarization of the homology matrix. The effect of the threshold will also be dependent on how closely related the target strains are to the reference strain. For example, within the strains in the Supplementary Tutorial notebooks we see that CU651637.1 and CP002167.1 are more dissimilar to reference model iML1515 as the drop off in retained genes occurs in a steeper fashion. We suggest using a threshold of 80% when comparing strains of the same species to ensure a sufficient similarity metric to include a gene in the draft models.

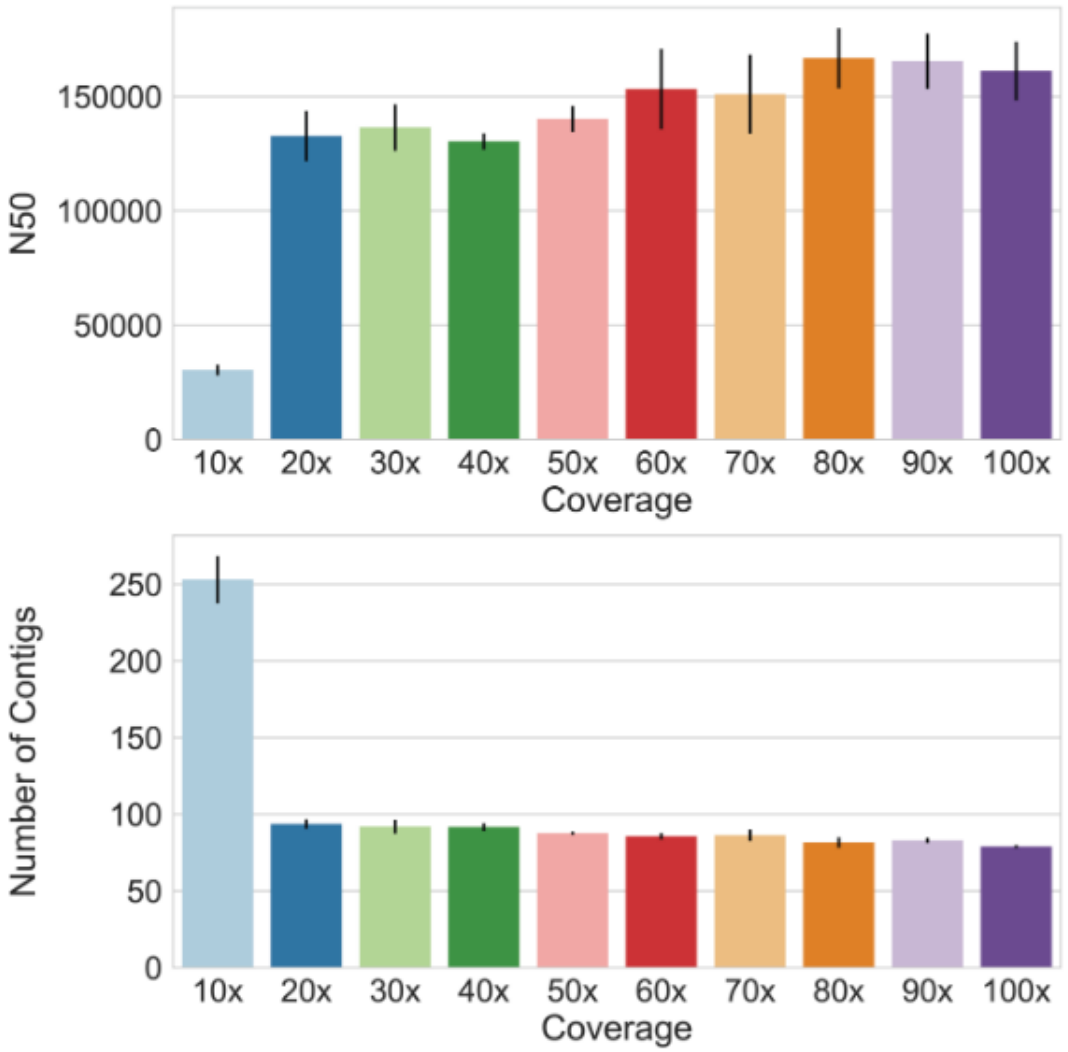

Figure D.2: to investigate the effect of coverage on overall assembly statistics of N50 and Number of Contigs, we randomly sampled reads of the BOP27 strain, which has been sequenced to extremely high coverage (400X), enabling this analysis. Analyzing the resulting assemblies at coverages ranging from 10X to 100X, we see from comparing the metrics that at 70X the assembly quality mostly saturates and as such we recommend included genomes have at least this much coverage

D.2 Supplementary Tutorial

The following section of this appendix contains the jupyter notebooks converted into pdf format found and referenced within Chapter 5: "A workflow for generating multi-strain genome-scale metabolic models of prokaryotes". This tutorial is best followed as interactive jupyter notebooks as intended with the original publication, however this static representation is included for reference.

1 Notebook 1: Homology matrix generation from genome sequences

In this tutorial, we will be working on generationg multi-strain genome-scale models for 5 E.coli strains. The reference model we used here is the iML1515 model published in Nature Biotechnology (PMID: 29020004), and the reference strain is E. coli K12 MG1655. We will be generating strain-specific models for 5 other E.coli strains: ATCC 8739, LF82, UM146, UMN026 and IAI39.

This is the the first notebook in the tutorial to create homology matrix from genome sequences.There are four major steps in this notebook 1. Download the genome annotation (GenBank files) from NCBI, and generate fasta files (protein &nucleotide) from them 2. Perform BLASTp to find homologous proteins in strains of interest 3. Use best bidirectional hits to create gene presence/absence matrix 4. Supplementary for best practice: use BLASTn to check if we have missed any unannotated open reading frames and retain these genes in orthology matrix as well as guide future manual curation

[1]:
```
#import packages needed
import pandas as pd
from glob import glob
from Bio import Entrez, SeqIO
```

[2]:
```
# Load the information on the five strains we will be working with in this␣
 ↪tutorial
StrainsOfInterest=pd.read_excel('Strain Information.xlsx')
StrainsOfInterest
```

[2]:

	Strain	NCBI ID	Pathotype
0	Escherichia coli ATCC 8739	CP000946.1	Commensal
1	Escherichia coli LF82	CU651637.1	InPec: AIEC
2	Escherichia coli UM146	CP002167.1	InPec: AIEC
3	Escherichia coli UMN026	CU928163.2	ExPec: UPEC
4	Escherichia coli IAI39	CU928164.2	ExPec: UPEC

[3]:
```
#The Reference Genome is as Described in the Base Reconstruction; in these␣
 ↪tutorials iML1515
referenceStrainID='NC_000913.3'
targetStrainIDs=list(StrainsOfInterest['NCBI ID'])
```

1.1 1. Download genome annotations (GenBank files) to generate fasta files

1.1.1 Dowload genomes from NCBI

Download the genome annotations (GenBank files) from NCBI for strains of interest.

[4]:
```
# define a function to download the annotated genebank files from NCBI
def dl_genome(id, folder='genomes'): # be sure get CORRECT ID
    files=glob('%s/*.gb'%folder)
    out_file = '%s/%s.gb'%(folder, id)
```

```
    if out_file in files:
        print (out_file, 'already downloaded')
        return
    else:
        print ('downloading %s from NCBI'%id)

    from Bio import Entrez
    Entrez.email = ""     #Insert email here for NCBI
    handle = Entrez.efetch(db="nucleotide", id=id, rettype="gb", retmode="text")
    fout = open(out_file,'w')
    fout.write(handle.read())
    fout.close()
```

[5]:
```
# execute the above function, and download the GenBank files for 5 E. coli␣
 ↪strains
for strain in targetStrainIDs:
    dl_genome(strain, folder='genomes')
```

```
downloading CP000946.1 from NCBI
downloading CU651637.1 from NCBI
downloading CP002167.1 from NCBI
downloading CU928163.2 from NCBI
downloading CU928164.2 from NCBI
```

1.1.2 Examine the Downloaded Strains

[6]:
```
# define a function to gather information of the downloaded strains from the␣
 ↪GenBank files
def get_strain_info(folder='genomes'):
    files = glob('%s/*.gb'%folder)
    strain_info = []

    for file in files:
        handle = open(file)
        record = SeqIO.read(handle, "genbank")

        for f in record.features:
            if f.type=='source':
                info = {}
                info['file'] = file
                info['id'] = file.split('\\')[-1].split('.')[0]
                for q in f.qualifiers.keys():
                    info[q] = '|'.join(f.qualifiers[q])
                strain_info.append(info)
    return pd.DataFrame(strain_info)
```

[7]:
```
# information on the downloaded strain
get_strain_info(folder='genomes')
```

[7]:
```
                  db_xref                    file                id  \
0  ATCC:8739|taxon:481805   genomes/CP000946.1.gb   genomes/CP000946
1           taxon:585056   genomes/CU928163.2.gb   genomes/CU928163
2           taxon:869729   genomes/CP002167.1.gb   genomes/CP002167
3           taxon:511145  genomes/NC_000913.3.gb  genomes/NC_000913
4           taxon:591946   genomes/CU651637.1.gb   genomes/CU651637
5           taxon:585057   genomes/CU928164.2.gb   genomes/CU928164

      mol_type                                   organism     strain  \
0  genomic DNA                Escherichia coli ATCC 8739  ATCC 8739
1  genomic DNA                   Escherichia coli UMN026     UMN026
2  genomic DNA                    Escherichia coli UM146      UM146
3  genomic DNA  Escherichia coli str. K-12 substr. MG1655       K-12
4  genomic DNA                     Escherichia coli LF82       LF82
5  genomic DNA                    Escherichia coli IAI39      IAI39

  sub_strain
0        NaN
1        NaN
2        NaN
3     MG1655
4        NaN
5        NaN
```

1.1.3 Generate FASTA files for both Protein and Nucleotide Pipelines

From the GenBank file, we can extract sequence and annoation information to generate fasta files for the protein and nucleotide analyses. The resulting fasta files will then be used in step 2 as input for BLAST

[8]:
```
# define a function to parse the Genbank file to generate fasta files for both␣
 ↪protein and nucleotide sequences
def parse_genome(id, type='prot', in_folder='genomes', out_folder='prots',␣
 ↪overwrite=1):

    in_file = '%s/%s.gb'%(in_folder, id)
    out_file='%s/%s.fa'%(out_folder, id)
    files =glob('%s/*.fa'%out_folder)

    if out_file in files and overwrite==0:
        print (out_file, 'already parsed')
        return
    else:
        print ('parsing %s'%id)
```

```
    handle = open(in_file)

    fout = open(out_file,'w')
    x = 0

    records = SeqIO.parse(handle, "genbank")
    for record in records:
        for f in record.features:
            if f.type=='CDS':
                seq=f.extract(record.seq)

                if type=='nucl':
                    seq=str(seq)
                else:
                    seq=str(seq.translate())

                if 'locus_tag' in f.qualifiers.keys():
                    locus = f.qualifiers['locus_tag'][0]
                elif 'gene' in f.qualifiers.keys():
                    locus = f.qualifiers['gene'][0]
                else:
                    locus = 'gene_%i'%x
                    x+=1
                fout.write('>%s\n%s\n'%(locus, seq))
    fout.close()
```

[9]:
```
# Generate fasta files for 5 strains of interest
for strain in targetStrainIDs:
    parse_genome(strain, type='prot', in_folder='genomes', out_folder='prots')
    parse_genome(strain, type='nucl', in_folder='genomes', out_folder='nucl')
```

```
parsing CP000946.1
parsing CP000946.1
parsing CU651637.1
parsing CU651637.1
parsing CP002167.1
parsing CP002167.1
parsing CU928163.2
parsing CU928163.2
parsing CU928164.2
parsing CU928164.2
```

[10]:
```
#Also generate fasta files for the reference strain
parse_genome(referenceStrainID, type='nucl', in_folder='genomes',␣
 ↪out_folder='nucl')
```

```
parse_genome(referenceStrainID, type='prots', in_folder='genomes',␣
 ↪out_folder='prots')
```

```
parsing NC_000913.3
parsing NC_000913.3
```

```
/home/cnorsig/.local/lib/python3.5/site-packages/Bio/Seq.py:2423:
BiopythonWarning: Partial codon, len(sequence) not a multiple of three.
Explicitly trim the sequence or add trailing N before translation. This may
become an error in future.
  BiopythonWarning)
```

1.2 2. Perform BLAST to find homologous proteins in strains of interest

1.2.1 Make BLAST DB for each of the target strains for both Protein and Nucleotide Pipelines

In this tutorial, we will run both BLASTp for proteins and BLSATn for nucleotides. BLASTp will be used as the main approach to identify homologous proteins in reference strain and other strains of interest, while BLASTn will be used as a supplementary method to check for any unannotated genes

[12]:
```
# Define a function to make blast database for either protein of nucleotide
def make_blast_db(id,folder='prots',db_type='prot'):
    import os

    out_file ='%s/%s.fa.pin'%(folder, id)
    files =glob('%s/*.fa.pin'%folder)

    if out_file in files:
        print (id, 'already has a blast db')
        return
    if db_type=='nucl':
        ext='fna'
    else:
        ext='fa'

    cmd_line='makeblastdb -in %s/%s.%s -dbtype %s' %(folder, id, ext, db_type)

    print ('making blast db with following command line...')
    print (cmd_line)
    os.system(cmd_line)
```

[13]:
```
# make protein sequence databases
# Because we are performing bi-directional blast, we make databases from both␣
 ↪reference strain and strains of interest
for strain in targetStrainIDs:
    make_blast_db(strain,folder='prots',db_type='prot')
make_blast_db(referenceStrainID,folder='prots',db_type='prot')
```

1.2.2 Define functions to run protein BLAST and get sequence lengths

- BLASTp will be the main approach used here to identify homologous proteins between strains
- Aside from sequence similarity, we also want to ensure the coverage of sequence mapping is sufficient. Therefore, we need to identiy the sequence length for each protein and compare it with the alignment length.

[15]:
```
# define a function to run BLASTp
def run_blastp(seq,db,in_folder='prots', out_folder='bbh',␣
↪out=None,outfmt=6,evalue=0.001,threads=1):
    import os
    if out==None:
        out='%s/%s_vs_%s.txt'%(out_folder, seq, db)
        print(out)

    files =glob('%s/*.txt'%out_folder)
    if out in files:
        print (seq, 'already blasted')
        return

    print ('blasting %s vs %s'%(seq, db))

    db = '%s/%s.fa'%(in_folder, db)
    seq = '%s/%s.fa'%(in_folder, seq)
    cmd_line='blastp -db %s -query %s -out %s -evalue %s -outfmt %s -num_threads␣
↪%i' \
    %(db, seq, out, evalue, outfmt, threads)

    print ('running blastp with following command line...')
    print (cmd_line)
    os.system(cmd_line)
    return out
```

[16]:
```
# define a function to get sequence length

def get_gene_lens(query, in_folder='prots'):

    file = '%s/%s.fa'%(in_folder, query)
    handle = open(file)
    records = SeqIO.parse(handle, "fasta")
    out = []

    for record in records:
        out.append({'gene':record.name, 'gene_length':len(record.seq)})

    out = pd.DataFrame(out)
    return out
```

1.3 3. Use Bi-Directional BLASTp Best Hits to create gene presence/absence matrix

1.3.1 Obtain Bi-Directional BLASTp Best Hits

From the above BLASTp results, we can obtain Bi-Directional BLASTp Best Hits to identify homologous proteins. Note beside gene similarity score, the coverage of alignment is also used to filter mapping results.

[17]:
```
# define a function to get Bi-Directional BLASTp Best Hits
def get_bbh(query, subject, in_folder='bbh'):

    #Utilize the defined protein BLAST function
    run_blastp(query, subject)
    run_blastp(subject, query)

    query_lengths = get_gene_lens(query, in_folder='prots')
    subject_lengths = get_gene_lens(subject, in_folder='prots')

    #Define the output file of this BLAST
    out_file = '%s/%s_vs_%s_parsed.csv'%(in_folder,query, subject)
    files=glob('%s/*_parsed.csv'%in_folder)

    #Combine the results of the protein BLAST into a dataframe
    print ('parsing BBHs for', query, subject)
    cols = ['gene', 'subject', 'PID', 'alnLength', 'mismatchCount',␣
 ↪'gapOpenCount', 'queryStart', 'queryEnd', 'subjectStart', 'subjectEnd',␣
 ↪'eVal', 'bitScore']
    bbh=pd.read_csv('%s/%s_vs_%s.txt'%(in_folder,query, subject), sep='\t',␣
 ↪names=cols)
    bbh = pd.merge(bbh, query_lengths)
    bbh['COV'] = bbh['alnLength']/bbh['gene_length']

    bbh2=pd.read_csv('%s/%s_vs_%s.txt'%(in_folder,subject, query), sep='\t',␣
 ↪names=cols)
    bbh2 = pd.merge(bbh2, subject_lengths)
    bbh2['COV'] = bbh2['alnLength']/bbh2['gene_length']
    out = pd.DataFrame()

    # Filter the genes based on coverage
    bbh = bbh[bbh.COV>=0.25]
    bbh2 = bbh2[bbh2.COV>=0.25]

    #Delineate the best hits from the BLAST
    for g in bbh.gene.unique():
        res = bbh[bbh.gene==g]
        if len(res)==0:
            continue
        best_hit = res.loc[res.PID.idxmax()]
```

```
        best_gene = best_hit.subject
        res2 = bbh2[bbh2.gene==best_gene]
        if len(res2)==0:
            continue
        best_hit2 = res2.loc[res2.PID.idxmax()]
        best_gene2 = best_hit2.subject
        if g==best_gene2:
            best_hit['BBH'] = '<=>'
        else:
            best_hit['BBH'] = '->'
        out=pd.concat([out, pd.DataFrame(best_hit).transpose()])

    #Save the final file to a designated CSV file
    out.to_csv(out_file)
```

[18]:
```
# Execute the BLAST for each target strain against the reference strain, save␣
 ↪results to 'bbh' i.e. "bidirectional best
# hits" folder to create
# homology matrix

for strain in targetStrainIDs:
    get_bbh(referenceStrainID,strain, in_folder='bbh')
```

1.3.2 Parse the BLAST Results into one Homology Matrix of the Reconstruction Genes

For the homology matrix, we only focus on genes that are present in the reference model

[19]:
```
#Load all the BLAST files between the reference strain and target strains

blast_files=glob('%s/*_parsed.csv'%'bbh')

for blast in blast_files:
    bbh=pd.read_csv(blast)
    print (blast,bbh.shape)
```

```
bbh/NC_000913.3_vs_CP002167.1_parsed.csv (3974, 16)
bbh/NC_000913.3_vs_CU651637.1_parsed.csv (3922, 16)
bbh/NC_000913.3_vs_CU928163.2_parsed.csv (4071, 16)
bbh/NC_000913.3_vs_CP000946.1_parsed.csv (4056, 16)
bbh/NC_000913.3_vs_CU928164.2_parsed.csv (3991, 16)
```

[20]:
```
#Load the base reconstruction to designate the list of genes within the model
import cobra
model = cobra.io.load_json_model('iML1515.json')
listGeneIDs=[]
for gene in model.genes:
    listGeneIDs.append(gene.id)
```

[21]:
```
#Create 2 matrices of N, rows where N is the number of model genes and M columns␣
 ↪where M is the number of target strains
#One matrix will be populated with the PID results from the blasts and another␣
 ↪with the mapping of gene locus tags

ortho_matrix=pd.DataFrame(index=listGeneIDs,columns=targetStrainIDs)
geneIDs_matrix=pd.DataFrame(index=listGeneIDs,columns=targetStrainIDs)
```

[22]:
```
#Parse through each blast file and acquire pertinent information for each matrix␣
 ↪for each of the base reconstruction genes
for blast in blast_files:
    bbh=pd.read_csv(blast)
    listIDs=[]
    listPID=[]
    for r,row in ortho_matrix.iterrows():
        try:
            currentOrtholog=bbh[bbh['gene']==r].reset_index()
            listIDs.append(currentOrtholog.iloc[0]['subject'])
            listPID.append(currentOrtholog.iloc[0]['PID'])
        except:
            listIDs.append('None')
            listPID.append(0)
    for col in ortho_matrix.columns:
        if col in blast:
            ortho_matrix[col]=listPID
            geneIDs_matrix[col]=listIDs
```

1.3.3 Apply Similarity Threshold to Binarize Homology Matrix to Presence/Absence Matrix

In this step, choose a threshold for the PID to determine if a gene is a absent/present in the strain of interest. We can then convert the homology matrix generated above into a binarized presence/absence matrix

[23]:
```
# In this tutoriao, genes with a greater than 80% PID are considered present in␣
 ↪the target strain genome
# and consequently less than 80% are considered absent from the target strain␣
 ↪genome
for column in ortho_matrix:
    ortho_matrix.loc[ortho_matrix[column]<=80.0,column]=0
    ortho_matrix.loc[ortho_matrix[column]>80.0,column]=1
```

1.4 4. Perform BLASTn to check unannotated open reading frames to guide manual curation

At this juncture it may be useful to execute a supplementary nucleotide BLAST to check for unannotated genes, results here become candidates for manual curation. In this tutorial we retain unannotated genes that pass the threhsold for similarity and contain no premature stop codons

[43]:
```
#Define a function to generate FNA from the GBK files
def gbk2fasta(gbk_filename):
    faa_filename = '.'.join(gbk_filename.split('.')[:-1])+'.fna'
    input_handle  = open(gbk_filename, "r")
    output_handle = open(faa_filename, "w")

    for seq_record in SeqIO.parse(input_handle, "genbank") :
        print ("Converting GenBank record %s" % seq_record.id)
        output_handle.write(">%s %s\n%s\n" % (
               seq_record.id,
               seq_record.description,
               seq_record.seq))

    output_handle.close()
    input_handle.close()
```

[44]:
```
#Define function to run the BLASTn
def run_blastn(seq, db,outfmt=6,evalue=0.001,threads=1):
    import os
    out = 'nucl/'+seq+'_vs_'+db+'.txt'
    seq = 'nucl/'+seq+'.fa'
    db = 'genomes/'+db+'.fna'

    cmd_line='blastn -db %s -query %s -out %s -evalue %s -outfmt %s -num_threads␣
 ↪%i' \
    %(db, seq, out, evalue, outfmt, threads)

    print ('running blastn with following command line...')
    print (cmd_line)
    os.system(cmd_line)
    return out
```

[42]:
```
# make nucleotide sequence databases
for strain in targetStrainIDs:
    make_blast_db(strain,folder='genomes',db_type='nucl')
```

[45]:
```
# convert genbank files to fna files for strains of interest
for strain in targetStrainIDs:
    gbk2fasta('genomes/'+strain+'.gb')
```

```
Converting GenBank record CP000946.1
Converting GenBank record CU651637.1
Converting GenBank record CP002167.1
Converting GenBank record CU928163.2
Converting GenBank record CU928164.2
```

[46]:
```
# perform uni-directional BLASTn hit
genome_blast_res=[]
for strain in targetStrainIDs:
    res = run_blastn(referenceStrainID,strain)
    genome_blast_res.append(res)
```

[47]:
```
#define a function to parse through the nucleotide BLAST results and form one␣
 ↪matrix of all the results
def parse_nucl_blast(infile):
    cols = ['gene', 'subject', 'PID', 'alnLength', 'mismatchCount',␣
 ↪'gapOpenCount', 'queryStart', 'queryEnd', 'subjectStart', 'subjectEnd',␣
 ↪'eVal', 'bitScore']
    data = pd.read_csv(infile, sep='\t', names=cols)
    data = data[(data['PID']>80) & (data['alnLength']>0.8*data['queryEnd'])]
    data2=data.groupby('gene').first()
    return data2.reset_index()
```

[48]:
```
# parse the nucleotide blast matrix
na_matrix=pd.DataFrame()
for file in genome_blast_res:
    genes =parse_nucl_blast(file)
    name ='.'.join(file.split('_')[-1].split('.')[:-1])
    na_matrix = na_matrix.append(genes[['gene','subject','PID']])
na_matrix = pd.pivot_table(na_matrix, index='gene',␣
 ↪columns='subject',values='PID')
```

[49]:
```
na_matrix.head()
```

[49]:

subject	CP000946.1	CP002167.1	CU651637.1	CU928163.2	CU928164.2
gene					
b0002	97.97	97.69	97.56	98.78	98.86
b0003	98.71	98.29	98.29	98.29	99.68
b0004	98.99	98.06	97.75	98.21	97.82
b0005	99.66	91.09	97.03	97.98	98.73
b0006	99.23	97.68	97.55	98.46	98.58

1.4.1 Examine unnannotated open reading frames

We compare the results from BLASTp and BLASTn and record any inconsistencies between the two matrices due to missing annotation. This result is then saved to guide future manual curation.

[50]:
```
# define a function to extract the sequence from fna file
def extract_seq(g, contig, start, end):
    from Bio import SeqIO
    handle = open(g)
    records = SeqIO.parse(handle, "fasta")
```

```
    for record in records:
        if record.name==contig:
            if end>start:
                section = record[start:end]
            else:
                section = record[end-1:start+1].reverse_complement()

            seq = str(section.seq)
    return seq
```

[51]:
```
#Define updated matrices that will include genes based on sequence evidence that␣
 ↪were missing due to lack of annotation
ortho_matrix_w_unannotated = ortho_matrix.copy()
geneIDs_matrix_w_unannotated = geneIDs_matrix.copy()
```

[52]:
```
#Define matrix of the BLASTn results for all the pertinent model genes
nonModelGenes=[]
for g in na_matrix.index:
    if g not in listGeneIDs:
        nonModelGenes.append(g)

na_model_genes=na_matrix.drop(nonModelGenes)
```

[53]:
```
#For each strain in the ortho_matrix, identify genes that meet threshold of SEQ␣
 ↪similarity, but missing from
#annotated ORFS. Additionally, look at the sequence to ensure that these cases␣
 ↪do not have early stop codons indicating
#nonfunctional even if the NA seqs meet the threshold

pseudogenes = {}

for c in ortho_matrix.columns:

    orfs = ortho_matrix[c]
    genes = na_model_genes[c]
    # All the Model Genes that met the BLASTp Requirements
    orfs2 = orfs[orfs==1].index.tolist()
    # All the Model Genes based off of BLASTn similarity above threshold of 80
    genes2 = genes[genes>=80].index.tolist()
    # By Definition find the genes that pass sequence threshold but were NOT in␣
 ↪annotated ORFs:
    unannotated = set(genes2) -set(orfs2)

    # Obtain sequences of this list to check for premature stop codons:
    data = 'nucl/NC_000913.3_vs_%s.txt'%c
```

```
    cols = ['gene', 'subject', 'PID', 'alnLength', 'mismatchCount',␣
↪'gapOpenCount', 'queryStart', 'queryEnd', 'subjectStart', 'subjectEnd',␣
↪'eVal', 'bitScore']
    data = pd.read_csv(data, sep='\t', names=cols)
    #
    pseudogenes[c] = {}
    unannotated_data = data[data['gene'].isin(list(unannotated))]
    for i in unannotated_data.index:
        gene = data.loc[i,'gene']
        contig = data.loc[i,'subject']
        start = data.loc[i,'subjectStart']
        end = data.loc[i,'subjectEnd']
        seq = extract_seq('genomes/%s.fna'%c,contig, start, end)
        # check for early stop codons - these are likely nonfunctional and␣
↪shouldn't be included
        if '*' in seq:
            print (seq)
            pseudogenes[c][gene]=seq
            # Remove the gene from list of unannotated genes
            unannotated-set([gene])


    print (c, unannotated)

    # For pertinent genes, retain those based off of nucleotide similarity␣
↪within the orthology matrix and geneIDs matrix
    ortho_matrix_w_unannotated.loc[unannotated,c]=1
    for g in unannotated:
        geneIDs_matrix_w_unannotated.loc[g,c] = '%s_ortholog'%g
```

```
CP000946.1 {'b4321', 'b0973', 'b0516', 'b3577', 'b1621', 'b1817', 'b1616',
'b2483', 'b0030', 'b4513', 'b1771'}
CU651637.1 {'b4321', 'b2930', 'b2344', 'b2690', 'b2430', 'b4513', 'b1588',
'b0150'}
CP002167.1 {'b4321', 'b2930', 'b4086', 'b1587', 'b0936', 'b2519', 'b2430',
'b3715', 'b4513', 'b3579', 'b1897'}
CU928163.2 {'b4513', 'b4515'}
CU928164.2 {'b4513', 'b4515'}
```

[54]:
```
#Save the Presence/Absence Matrix and geneIDs Matrix for future use
ortho_matrix_w_unannotated.to_csv('ortho_matrix.csv')
geneIDs_matrix_w_unannotated.to_csv('geneIDs_matrix.csv')
```

2 Notebook 2: Generate multi-strain models

This notebook follows Notebook 1 in the tutorial, and continues to work on generating multi-strain E.coli models. This notebook utilizes the output of notebook 1 (presence/absence matrix and geneID matrix) to generate draft strain-specific models from the reference model. There are two major steps involved:

1. Deletion of missing genes/reaction from reference model to generate draft models
2. Update gene-protein-reaction rule in each model

[4]:
```
#import package needed
import cobra
import pandas as pd
from cobra.io import load_json_model
from glob import glob
from cobra.manipulation.delete import delete_model_genes, remove_genes
```

2.1 1. Deletion of missing genes/reaction from reference model

[5]:
```
#Load the base E. coli reconstruction iML1515
model = load_json_model('iML1515.json')
model
```

[5]:
```
<Model iML1515 at 0x7f8929212940>
```

[6]:
```
## Load the previously generated homology matrix for E. coli strains of interest
hom_matrix=pd.read_csv('ortho_matrix.csv')
hom_matrix=hom_matrix.set_index('Unnamed: 0')
```

2.1.1 Delete missing genes from copies of the iML1515 model

For each strain, start with the iML1515 model, identify the missing genes from the matrix, and remove them and their associated reactions from the reference model

[5]:
```
#create strain-specific draft models and save them
for strain in hom_matrix.columns:

    #Get the list of Gene IDs from the homology matrix dataframe for the current␣
 ↪strain without a homolog
    currentStrain=hom_matrix[strain]
    nonHomologous=currentStrain[currentStrain==0.0]
    nonHomologous=nonHomologous.index.tolist()

    #s0001 is an artificial gene used in iML1515 for spontaneous reactions and␣
 ↪as such has no homologs,
    #However, it is retained for these spontaneous reactions to function
    nonHomologous.remove('s0001')
```

```
    #Define a list of Gene objects from the base reconstruction to be deleted␣
↪from the current strain
    toDelete=[]
    for gene in nonHomologous:
        toDelete.append(model.genes.get_by_id(gene))

    #Establish a model copy and use the COBRApy function to remove the␣
↪appropriate content and save this model
    modelCopy=model.copy()
    remove_genes(modelCopy, toDelete, remove_reactions=True)
    modelCopy.id=str(strain)
    cobra.io.json.save_json_model(modelCopy, str('Models/'+strain+'.json'),␣
↪pretty=False)
```

2.2 2. Update Model Gene Product Rules

[ ]:
```
#load the geneID matrix from the notebook1
models=glob('%s/*.json'%'Models')
geneIDs_matrix=pd.read_csv('geneIDs_matrix.csv')
geneIDs_matrix=geneIDs_matrix.set_index('Unnamed: 0')
```

[8]:
```
#Utilize the geneIDs matrix to update the GPRs in each of the strain-specific␣
↪models with the proper gene ID

from cobra.manipulation.modify import rename_genes

for mod in models:
    model=cobra.io.load_json_model(mod)
    for column in geneIDs_matrix.columns:
        if column in mod:
            currentStrain=column

    IDMapping=geneIDs_matrix[currentStrain].to_dict()
    IDMappingParsed = {k:v for k,v in IDMapping.items() if v != 'None'}

    rename_genes(model,IDMappingParsed)
    cobra.io.json.save_json_model(model,mod, pretty=False)
```

2.2.1 Examine the draft strain specific model contents

[9]:
```
# gather the general information on the draft models
for strain in hom_matrix.columns:
    model=cobra.io.load_json_model(str('Models/'+strain+'.json'))
    print (model.id,'Number of Model Genes:',len(model.genes),'Number of Model␣
↪Reactions:',len(model.reactions))
```

```
CP000946.1 Number of Model Genes: 1484 Number of Model Reactions: 2664
CU651637.1 Number of Model Genes: 1420 Number of Model Reactions: 2613
CP002167.1 Number of Model Genes: 1425 Number of Model Reactions: 2614
CU928163.2 Number of Model Genes: 1460 Number of Model Reactions: 2650
CU928164.2 Number of Model Genes: 1459 Number of Model Reactions: 2622
```

3 Notebook 3: Investigate Strain-Specific Capabilities

In this notebook, we showcase some simple applications of strain-specific models including growth simulation on different media. This notebook utilize the draft models created from Notebook 2. Note that the models constructed and used here could undergo further manual curation to increase their quality and content.

[3]:
```
import cobra
import pandas as pd
import seaborn as sns
from cobra.io import load_json_model
from glob import glob
```

[4]:
```
#load the draft models created from Notebook2
model_files=glob('%s/*.json'%'Models')
model_files
```

[4]:
```
['Models/CP000946.1.json',
 'Models/CU651637.1.json',
 'Models/CU928164.2.json',
 'Models/CP002167.1.json',
 'Models/CU928163.2.json']
```

3.1 Begin Constraint-Based Modeling on Group of Strain-Specific Models

[5]:
```
#Establish a definition that initializes models to an in silico representation␣
 ↪of M9 media

def m9(model):
    for reaction in model.reactions:
        if 'EX_' in  reaction.id:
            reaction.lower_bound=0

    model.reactions.EX_ca2_e.lower_bound=-1000
    model.reactions.EX_cl_e.lower_bound=-1000
    model.reactions.EX_co2_e.lower_bound=-1000
    model.reactions.EX_cobalt2_e.lower_bound=-1000
    model.reactions.EX_cu2_e.lower_bound=-1000
    model.reactions.EX_fe2_e.lower_bound=-1000
    model.reactions.EX_fe3_e.lower_bound=-1000
```

```
    model.reactions.EX_h_e.lower_bound=-1000
    model.reactions.EX_h2o_e.lower_bound=-1000
    model.reactions.EX_k_e.lower_bound=-1000
    model.reactions.EX_mg2_e.lower_bound=-1000
    model.reactions.EX_mn2_e.lower_bound=-1000
    model.reactions.EX_mobd_e.lower_bound=-1000
    model.reactions.EX_na1_e.lower_bound=-1000
    model.reactions.EX_tungs_e.lower_bound=-1000
    model.reactions.EX_zn2_e.lower_bound=-1000
    model.reactions.EX_ni2_e.lower_bound=-1000
    model.reactions.EX_sel_e.lower_bound=-1000
    model.reactions.EX_slnt_e.lower_bound=-1000
    model.reactions.EX_glc__D_e.lower_bound=-20
    model.reactions.EX_so4_e.lower_bound=-1000
    model.reactions.EX_nh4_e.lower_bound=-1000
    model.reactions.EX_pi_e.lower_bound=-1000
    model.reactions.EX_cbl1_e.lower_bound=-.01
    model.reactions.EX_o2_e.lower_bound=-20

    return model
```

[6]:
```
#Load each target Strain model, initialize it to glucose M9 media and see if the␣
 ↪model can optimize for
# biomass production

for model in model_files:
    mod=cobra.io.load_json_model(model)
    m9(mod)
    print (mod.id, mod.optimize().f)
```

```
CP000946.1 1.1207961081933524
CU651637.1 1.1207855410255514
CU928164.2 1.1207855410255527
CP002167.1 1.1207855410255525
CU928163.2 1.1207961081933542
```

[7]:
```
#In this Tutorial we see that all of the target-strain models are immediately␣
 ↪able to solve in the defined medium
#This will not always be the case and gap-filling and identification of␣
 ↪auxotrophies may be necessary
#(see original protocol)
```

3.2 Example of examining strain-specific capabilities: carbon source utilization

In this example, we examine the draft models abilities to simulate growth on different carbon sources. The carbon sources are limited to those with exchange reactions in the model

[8]:
```
# load the reference model and extract the list of carbon source to test

model = load_json_model('iML1515.json')

StrainsOfInterest=pd.read_excel('Strain Information.xlsx')
targetStrainIDs=list(StrainsOfInterest['NCBI ID'])

listPotCarbonSources=[]
for r in model.reactions:
    if 'EX_' in r.id:
        for m in r.metabolites:
            if 'C' in m.formula:
                listPotCarbonSources.append(r.id)
```

[9]:
```
#create a dataframe to store the simulation result on list of carbon source
growthCapabilities=pd.
 ↪DataFrame(index=listPotCarbonSources,columns=targetStrainIDs)
```

[10]:
```
#iterature through all the models to simulate growth on different carbon sources
# to do so, we closed the default carbon source glucos by setting the lower␣
 ↪bound of its exchange reaction to 0.
# and open the exchange reaction of the carbon source of interest to enable␣
 ↪nutrient update

for model in model_files:
    mod=cobra.io.load_json_model(model)
    listCapabilities=[]

    for source in listPotCarbonSources:
        m9(mod)
        mod.reactions.EX_glc__D_e.lower_bound=0
        mod.reactions.get_by_id(source).lower_bound=-1000
        listCapabilities.append(mod.optimize().f)

    for col in growthCapabilities.columns:
        if col in model:
            growthCapabilities[col]=listCapabilities
```

[11]:
```
growthCapabilities
```

[11]:

	CP000946.1	CU651637.1	CP002167.1	CU928163.2	CU928164.2
EX_acgam_e	12.842831	12.842748	12.842748	12.842831	12.842748
EX_cellb_e	11.878912	11.878870	11.878870	11.878912	11.878870
EX_chol_e	0.000000	0.221636	0.221636	0.000000	0.000000
EX_ade_e	0.496634	0.496627	0.496627	0.496634	0.281315
EX_4abut_e	0.562435	0.562428	0.562428	0.562435	0.562428
EX_ac_e	0.428358	0.428352	0.428352	0.428358	0.428352

EX_akg_e	8.364325	8.364251	8.364251	8.364325	8.364251
EX_ala__L_e	2.899720	2.899686	2.899686	2.899720	2.899686
EX_arg__L_e	2.798891	2.798856	2.798856	2.601610	2.798856
EX_asp__L_e	9.704300	9.704202	9.704202	9.704300	9.704202
EX_cytd_e	19.192908	19.192791	19.192791	19.192908	19.192791
EX_dcyt_e	18.878871	18.878697	18.878697	18.878871	18.878697
EX_fum_e	9.704300	9.704202	9.704202	9.704300	9.704202
EX_glu__L_e	7.951270	7.951201	7.951201	7.951270	7.951201
EX_gua_e	1.396844	1.396771	1.396771	1.396844	0.934816
EX_met__L_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_metsox_S__L_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_crn_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_glcn_e	16.301036	16.300949	16.300949	16.301036	16.300949
EX_gln__L_e	5.138213	5.138151	5.138151	5.138213	5.138151
EX_glyc_e	7.683620	7.683557	7.683557	7.683620	7.683557
EX_man_e	11.878912	11.878870	11.878870	11.878912	11.878870
EX_rib__D_e	5.746865	5.746832	5.746832	5.746865	5.746832
EX_sbt__D_e	9.614060	9.614019	9.614019	9.614060	9.614019
EX_ura_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_val__L_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_xan_e	0.481995	0.481989	0.481989	0.481995	0.256790
EX_co2_e	-0.043467	-0.043463	-0.043463	-0.043467	-0.043730
EX_hxan_e	0.494887	0.494880	0.494880	0.494887	0.280753
EX_ile__L_e	-0.043467	-0.043463	-0.043463	-0.043467	-0.043730
...	...	...	...	...	...
EX_urate_e	0.393785	0.393779	0.393779	0.393785	0.208581
EX_cpgn_un_e	-0.043467	-0.043463	-0.043463	-0.043467	-0.043730
EX_tartr__D_e	10.041344	10.041214	10.041214	10.041344	10.041214
EX_crn__D_e	0.000000	0.000000	0.000000	0.000000	0.000000
EX_psclys_e	10.165027	0.000000	0.000000	0.000000	0.000000
EX_galctn__L_e	16.005483	0.000000	0.000000	16.005483	0.000000
EX_5dglcn_e	0.000000	18.987283	18.987283	18.987322	18.987283
EX_ppal_e	0.428358	0.428352	0.428352	0.428358	0.428352
EX_LalaDglu_e	6.970883	6.970794	6.970794	6.970883	6.970794
EX_LalaLglu_e	6.970883	6.970794	6.970794	6.970883	6.970794
EX_ttrcyc_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_mincyc_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_doxrbcn_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_fusa_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_cm_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_novbcn_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_rfamp_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_quin_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_3hpp_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_5mtr_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_arbt_e	0.000000	-0.043463	0.000000	0.000000	0.000000
EX_dxylnt_e	0.000000	-0.043463	0.000000	6.285708	0.000000

```
EX_mththf_e          0.000000   -0.043463    0.000000    0.000000    0.000000
EX_dhps_e            0.000000   -0.043463    0.000000    0.000000    0.000000
EX_cs1_e             0.000000   -0.043463    0.000000    0.000000    0.000000
EX_mepn_e            0.000000   -0.043463    0.000000    0.000000    0.000000
EX_sq_e              0.000000   -0.043463    0.000000   14.547635   14.547530
EX_4abzglu_e         0.000000   -0.043463    0.000000    0.000000    0.000000
EX_metglcur_e       18.428009   18.427911   18.427911   18.428009   18.427911
EX_2dglc_e           0.000000    0.000000    0.000000    0.000000    0.000000

[298 rows x 5 columns]
```

[12]:
```
#use heatmap to visualize the difference in growth simulation on carbon sources
##Already differences in growth capabilities are apparent between just these 5␣
 ↪strains
sns.clustermap(growthCapabilities)
```

[12]:
```
<seaborn.matrix.ClusterGrid at 0x7f4ec11877f0>
```

3.3 Example of gap-filling a model through a custom gap-filling implementation defined below

In this example, we examine the draft model of CP000946.1 and it's inability to grow using EX_metglcur_e as a carbon source as this is readily apparent and different from the other draft models from the proceeding analysis

[ ]:
```
#Gather the list of base strain genes that have no homolog in strain of␣
 ↪interest, an input to the below function
hom_matrix=pd.read_csv('ortho_matrix.csv')
hom_matrix=hom_matrix.set_index('Unnamed: 0')
strain=hom_matrix['CP000946.1']
missingGenes=list(strain[strain==0.0].index)
```

[ ]:
```
def gapfill_multi(model, missing_genes, **kwargs):
    """
    Generate a list of gapfilling reactions from a list of missing genes for a␣
 ↪strain-specific model.

    :param model: COBRA model for the base strain with the objective coefficient␣
 ↪for the reaction of interest (e.g. biomass reaction) set to 1.

    :param missing_genes: list of genes with no homologs in the strain of␣
 ↪interest.

    :param lower_bound: minimum allowable yield of gapfilled model.

    :param biomass: override the current model settings and temporarily assign␣
 ↪the objective coefficient for a function of interest to 1.
```

```
    :return: a list of gapfilling reactions.

    """

    if 'lower_bound' in kwargs.keys():
        lower_bound = kwargs['lower_bound']
    else:
        lower_bound = model.optimize().objective_value*0.5

    biomass_reactions = [rx.id for rx in model.reactions if rx.
↪objective_coefficient == 1]
    if 'biomass' in kwargs.keys():
        biomass = kwargs['biomass']
        if len(biomass_reactions) > 1:
            for rx in set(biomass_reactions) - {biomass}:
                model.reactions.get_by_id(rx).objective_coefficient = 0

    else:
        if len(biomass_reactions) > 1:
            raise Exception("This model has more than one objective. \n Please␣
↪adjust the objective coefficient to 1 for the chosen objective reaction (e.g.␣
↪biomass or ATP) and 0 for the rest of the reactions, \n or specify the␣
↪reaction ID to use as an objective.")
        if len(biomass_reactions) > 1:
            raise Exception("The model doesn't have an objective function.␣
↪Please set the appropriate objective coefficient to 1, or specify the reaction␣
↪ID to use as an objective.")
        biomass = biomass_reactions[0]


    model.solver.configuration.tolerances.feasibility = 1e-9
    constraints = []
    indicators = []

    for rx in cobra.manipulation.find_gene_knockout_reactions(model,␣
↪missing_genes):

        indicator = model.problem.Variable('%s_i'%rx.id , type = 'binary')
        indicators.append(indicator)

        new_cstr1 = model.problem.Constraint( rx.flux_expression - rx.
↪upper_bound*indicator ,ub = 0)
        new_cstr2 = model.problem.Constraint(-rx.flux_expression + rx.
↪lower_bound*indicator ,ub = 0)
        constraints += [new_cstr1, new_cstr2]
```

```
        model.add_cons_vars([new_cstr1, new_cstr2, indicator])

    model.reactions.get_by_id(biomass).lower_bound = lower_bound
    model.objective = model.problem.Objective(-sum(indicators))
    sol = model.optimize()
    indicator_results = [ind.name[:-2] for ind in indicators if ind.primal != 0.
↪0]

    # removing changes to model
    model.remove_cons_vars(constraints+indicators)
    for rx in set(biomass_reactions):
        model.reactions.get_by_id(rx).objective_coefficient = 1

    return indicator_results
```

[ ]:
```
# We see that in this condition the model cannot synthesize biomass in this␣
↪condition as per
# above analysis
model=cobra.io.load_json_model('Models/CP000946.1.json')
m9(model)
model.reactions.EX_glc__D_e.lower_bound=0
model.reactions.EX_metglcur_e.lower_bound=-1000
model.optimize()
```

[ ]:
```
# We see that however the base model can synthesize biomass in this condition
base=cobra.io.load_json_model('iML1515.json')
m9(base)
base.reactions.EX_glc__D_e.lower_bound=0
base.reactions.EX_metglcur_e.lower_bound=-1000
base.optimize()
```

[ ]:
```
#By runnning the above function we obtain the list of candidate reactions
gapfill_multi(base, missingGenes)
```

[ ]:
```
base.reactions.METGLCURt2pp.genes
```

[ ]:
```
#Upon further inspection we see that the lack of a homolog in the b1616 gene is␣
↪what causes the CP000946.1 strain
#to lose this functionality
'b1616' in missingGenes
```

Appendix E

Systems biology approach to functionally assess the *Clostridioides difficile* pan-genome reveals genetic diversity with discriminatory power

E.1 Supplementary Text

E.1.1 Functional Annotation Specific Driven Typing

In addition to using the full accessory genome to define strain groupings, we were interested in evaluating each functional annotation subset ability to drive typings. The full set of accessory genes was split into the 4 categories of Metabolism, Information Storage and Processing, Cellular Processes and Signaling, and Uncharacterized as described within the section "Characterizing the *C. difficile* pan-genome". Each of these 4 sets were then independently used following the same algorithm described when analyzing the full accessory genome. When using these subsets the 451 strains can be grouped into 94, 166, 182, and 209 groups based on accessory metabolic, information related, signaling related, and uncharacterized genes respectively. Grouping based on these subsets allows for insight into the differential similarity as exhibited by the RT002 strains which did not group into a singular all accessory content, but were exclusively grouped by only the metabolic related accessory genes.

E.1.2 Expanded Comparison to Additional Typing Schemes

In addition to our presented comparisons to the most-widely used MLST and PCR Ribotyping schemes, we also compared STAG to other available typing schemes that consider increased genetic content. We were able to establish groups through SNP-based, KMER-based, and a core-genome MLST (CGMLST) approach. Additionally, we also analyzed the ability of standard distance-based clustering methods, namely hierarchical clustering, to derive strain groups on the same accessory genome matrix used as input to STAG. These comparisons serve to evaluate two key aspects of the STAG approach: 1) the effect of increased amount of genetic content used to establish groups, and 2) the iterative sorting based on shifting similarity thresholds to establish

groups versus an existing clustering technique.

Overall, the SNP and CGMLST strain types had similar compression to the STAG PGTs, however a slightly greater percentage (59.6%, 60.4% respectively versus 55.6%) of the final strain types in each case were single strain sized (Supplementary Figure E.9). The KMER approach failed to establish meaningful groups and only compressed the 451 strains into 312 groups. Conversely, the hierarchical clustering based groupings resulted in 99 strain groups compared to the 176 PGTs with 48.5% of these groups being single strain groups. On average each typing scheme of STAG, CGMLST, KMER, and SNP resulted in comparable average strain group sizes of 2.56, 2.62, 1.44, 3.09 respectively. The use of hierarchical clustering necessitated the specification of a single distance threshold on which to define groups (.05 here) and the lack of flexibility in distance metrics leads to larger groups encapsulating more diverse strains more similarly to the MLST system. The largest group of strains when using hierarchical clustering is 68 strains, which results in limited interpretability. The goal of any strain-typing scheme is to provide a means to describe the variation of strains into meaningful groups. To this end each approach has its strengths and careful consideration of what content drives the resulting strain types is important. STAG presents another valuable tool that is easy and efficient to use and offers a high degree of interpretability and information in regard to accessory genome content.

E.2 Supplementary Figures

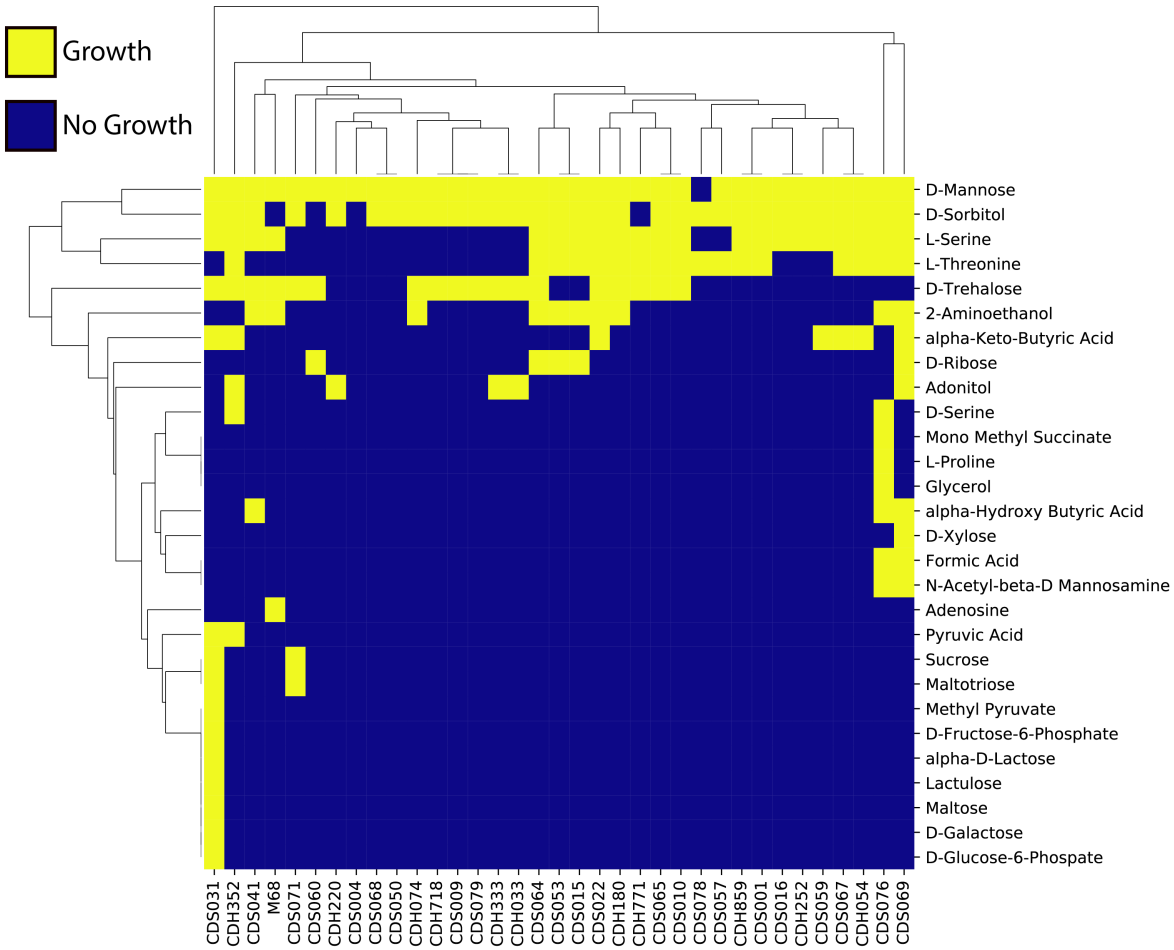

Figure E.1: 28 Non-Unanimous Growth Supporting Carbon Sources. The carbon sources where strains varied in terms of growth capabilities and as a result define the varying metabolic profiles.

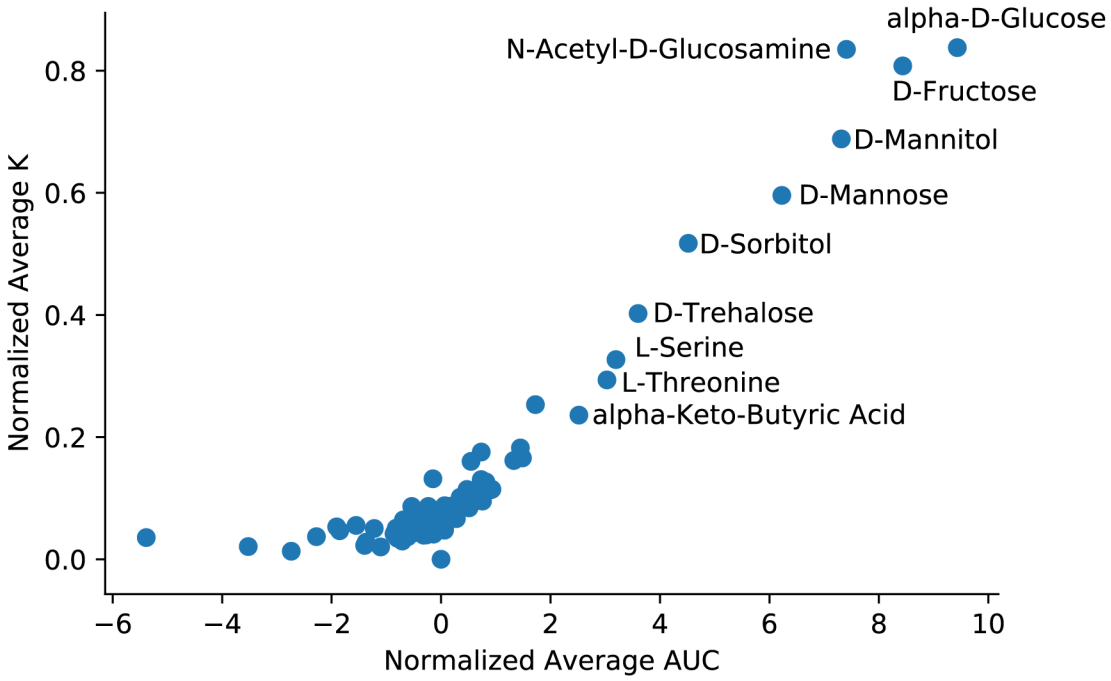

Figure E.2: Average fit parameters across 35 isolates. After averaging the AUC and K, the most growth supporting nutrients across our dataset emerge.

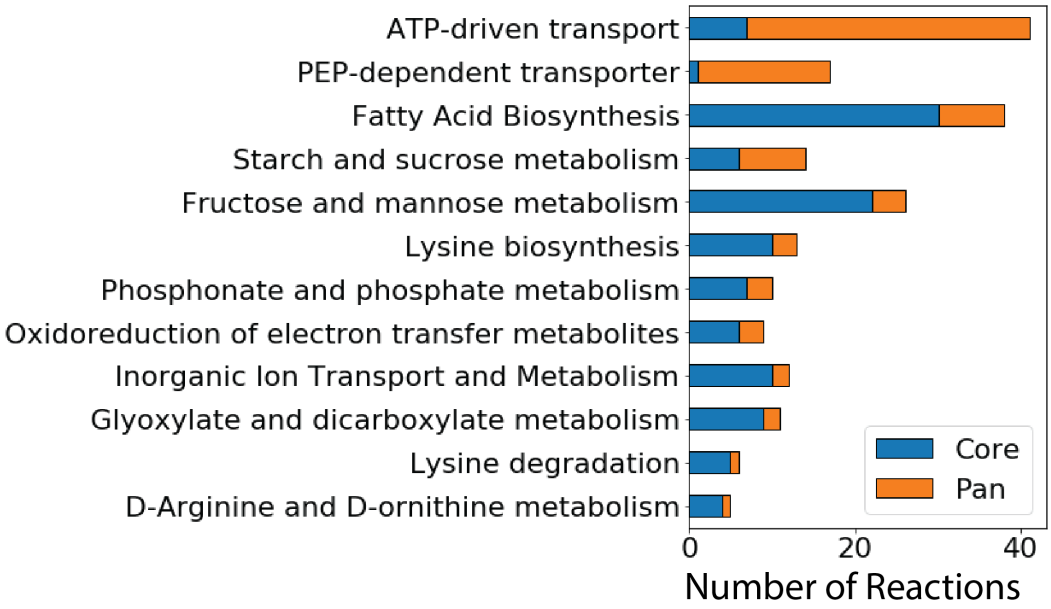

Figure E.3: Reaction Subsystems with High Degree of Non Conserved Genes. Pictured are the subsystems identified through the use of GEMs where greater than 15% of the reactions in the subsystem contain at least one non conserved gene within the set of 35 clinical isolates.

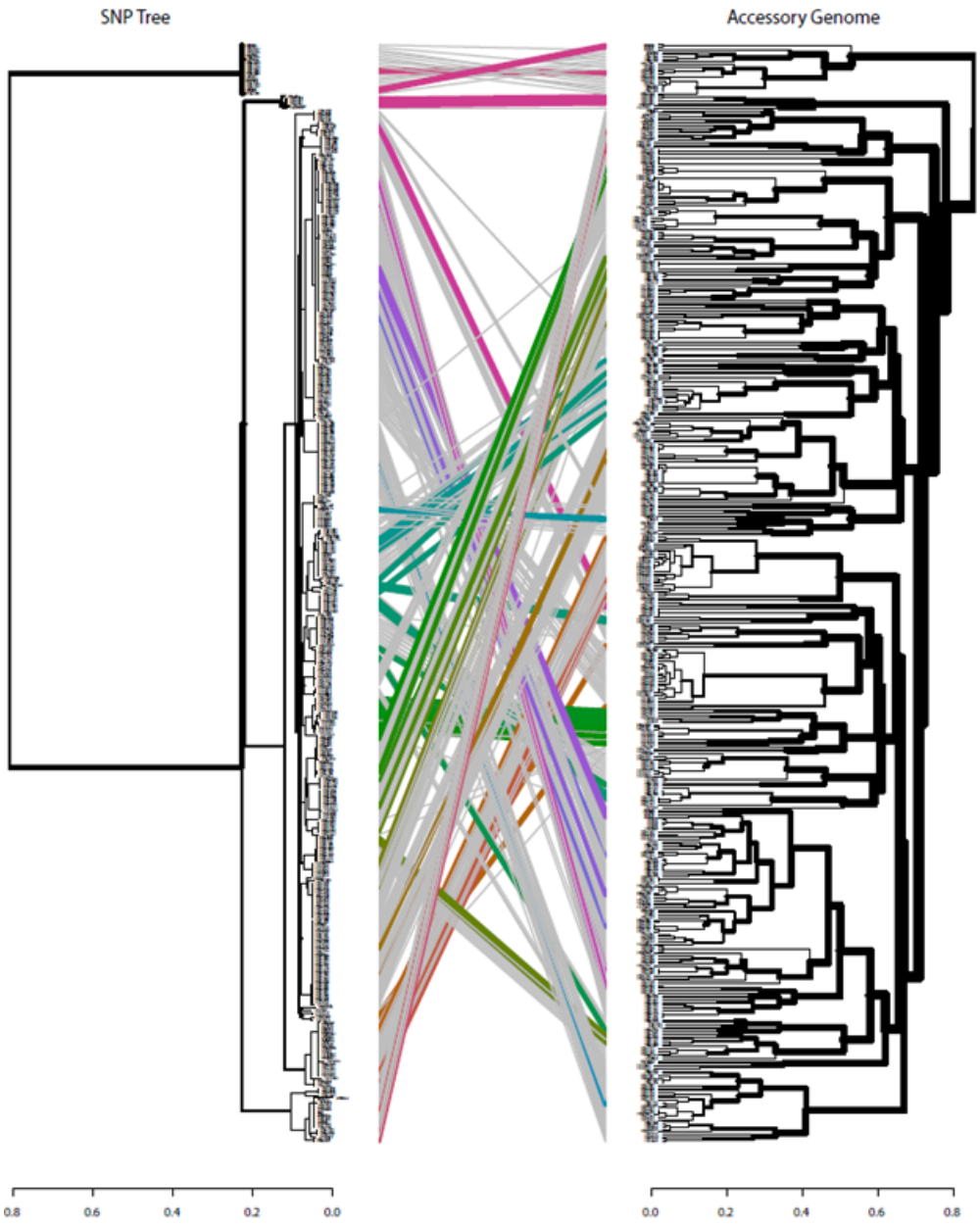

Figure E.4: Comparison of SNP-based dendrogram (left) and accessory-genome based dendrogram (right). Colored lines connecting the two trees indicate groups of strains with identical clustering hierarchies, gray lines indicate different clustering hierarchies. The two trees have a correlation of 0.55 and entanglement of 0.12 indicating that accessory genome content is not completely concordant with SNP-based phylogeny. Tree comparisons and visualizations were constructed using the dendextend package (see Methods).

Figure E.5: Accessory gene cluster presence absence for each strain. Pictured is each of the 451 strains (ordered by eventual PGT) and presence/absence of each of the 4,057 accessory gene clusters.

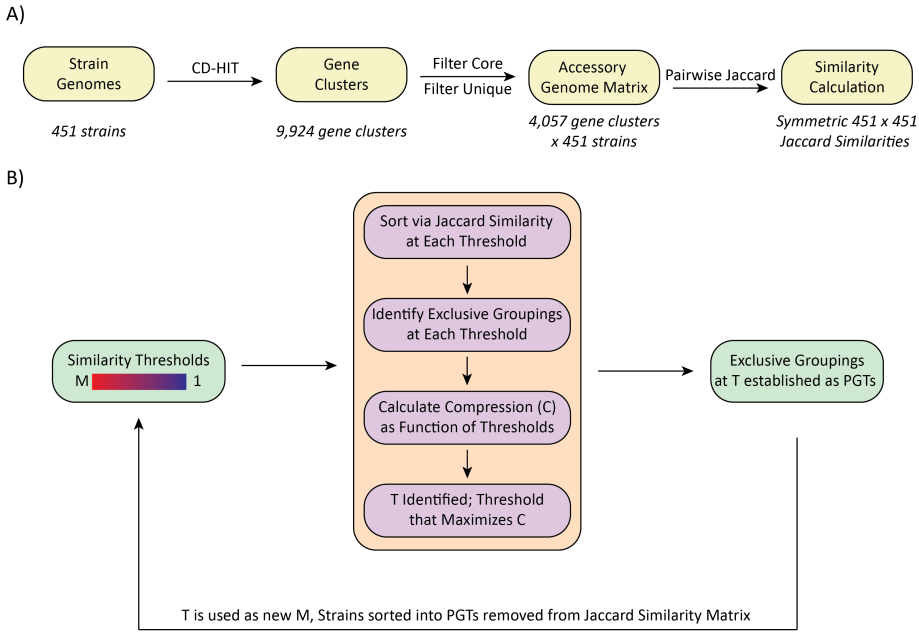

Figure E.6: STAG Workflow for Establishing Pan-Genome Typings. A) Strain genomes are first clustered using CD-Hit to establish the accessory genome. The Jaccard Similarity was calculated on accessory genome vectors between strains. B) A range of similarity thresholds are evaluated at each iterative pass of the sorting algorithm over the jaccard similarity matrix. Exclusive groupings are identified at each potential threshold and the threshold that maximized compression of exclusive groups is selected. Exclusive groups become PGTs and the identified threshold is the beginning threshold range, M, of the next pass. The strains sorted into an exclusive PGT are dropped from the similarity matrix.

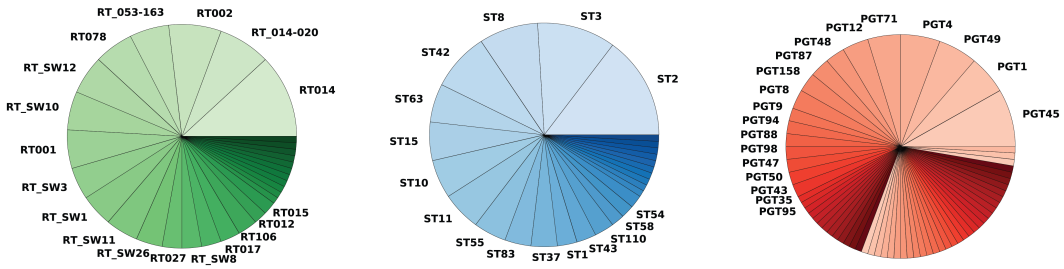

Figure E.7: Assigned Typings for All Strains with Ribotyping Information. For the 108 strains for which a typing is available in all three schemes, the relative composition of typings is shown. While there are strains of unique RT and MLST, there are a far greater number of strains that are assigned as unique PGT.

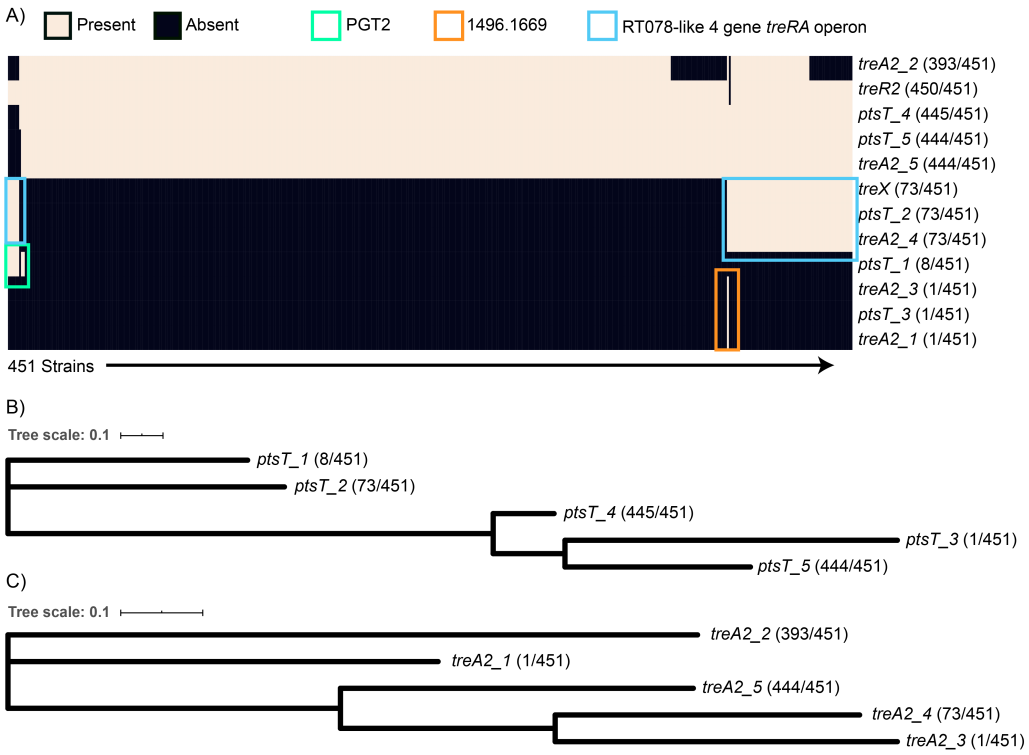

Figure E.8: Analysis of *treRA* operon characteristic to RT078 and related gene clusters. A) We identified the gene clusters within the pan-genome that include the sequences of a previously identified four-gene insertion responsible for increased ability to catabolize trehalose by RT078 strains. These clusters are present within 73 strains studied, including all strains of PGT1 which includes all RT078 strains within the study. The remainder of the strains represent a variety of Clade 1 strains that have also been shown to potentially include this gene insertion. By analyzing the identified related clusters to the *treA2* and *ptsT* gene clusters of the insertion we were able to identify variants unique to one strain as well as a variant *ptsT* specific to strains classified as PGT2. B) The alignment of the representative sequences for *ptsT* clusters demonstrates that the PGT2 specific variant is of closest relation to the known *ptsT* indicative of RT078. C) The alignment of the representative sequences for *treA2* clusters demonstrates that of the two unique variants to strain 1496.1669 one is more related to that within the RT078 operon and the other to genes more core genes in the population.

A

Dataset (Strains)	RTs	MLSTs	PGTs	CGMLST	KMER	SNP	Hierarchical
Clinical Isolates (35)	11	12	22	16	32	17	14
Public Strains with Annotated Ribotypes (108)	30	27	63	40	98	60	36
All Public Strains (451)	N/A	57	176	172	312	146	99

B

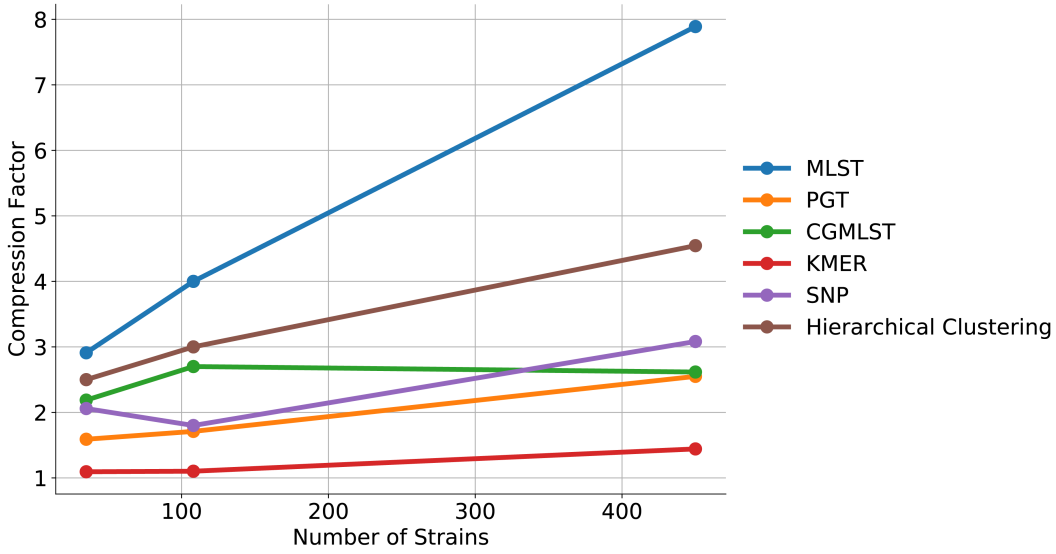

Figure E.9: Expanded comparison of different strain typing schemes. A) Number of strain groups identified as detailed within Figure 4 expanded to include CGMLST, KMER, SNP, and hierarchical clustering of the accessory genome. B) The compression factor as a function as a function of the number of strains typed demonstrates that the use of a single distance metric in hierarchical clustering results in groups in between MLST and CGMLST, SNP, and STAG in terms of resolution.