UNIVERSITY OF CALIFORNIA, SAN DIEGO

Comparative Systems Biology Analysis of Microbial Pathogens

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in

Chemical Engineering

by

Jonathan Mayock Monk

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Michael Heller, Co-Chair
Professor Guarav Arya
Professor Victor Nizet
Professor Milton Saier

2015

The dissertation of Jonathan Mayock Monk is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co- Chair

_____

Chair

University of California, San Diego

2015

DEDICATION

This work is dedicated to my parents, for the love and support they have given me that has enabled me to achieve this milestone.

EPIGRAPH

The true delight is in the finding out rather than in the knowing.

- Isaac Asimov

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many people have influenced me and deserve enormous thanks for helping me to get to this point. First, my advisor, Bernhard Palsson, is responsible for helping me become the scientist I am today. His inquisitive mind, constant encouragement and ample support have enabled me to achieve so much more in my PhD than I ever thought possible. No matter the problems we encountered along the way, he always found ways to re-frame the problem and new exciting angles of attack. Moreover, it has been my privilege to be part of the Systems Biology Research Group. Dr. Palsson has cultivated an open and dynamic environment that encourages students to interact, collaborate and share their ideas. The lab retreats always fostered strategic thinking and critical evaluation of our own capabilities. I learned that this self-reflection is vital to being an effective scientist. I also thank the members of my thesis committee, who helped to guide this dissertation along the way into the work it is today.

Adam Feist and Pep Charusanti also served as mentors to me during my time here. Pep taught me how to choose good scientific problems and how to write scientifically. Adam helped me learn to prioritize in graduate school. Both of them taught me about formulating a hypothesis, how to prove (or disprove!) that hypothesis and finally, the invaluable skill of how to communicate this process and its results to an audience.

I was lucky to have fabulous undergraduate mentors, including Princeton Professors Ron Weiss, Rick Register and Jay Benziger. Also, Bob Scogna, a graduate student, made graduate school look fun, rewarding and a goal worth pursuing. I also thank my high-school AP Biology teacher, Bob Doltar, who first got me excited about

biotechnology. Additionally, I must thank the funding sources that supported me financially to pursue the research presented in this thesis. The National Institutes of Health and Novo Nordisk Foundation provided the majority of my support throughout graduate school. The National Science Foundation and the Japanese Society for the Promotion of Science funded my three-month trip to Japan to research at the Nara Advanced Institute for Science and Technology, a fantastic experience.

I also thank my fellow graduate school lab-mates. First, Joshua Lerman, who introduced me to the SBRG by TAing me in BENG212. He taught me a solid fundamental basis of constraints-based modelling and served as a constant source of new and out-of-the-box ideas. Aarash Bordbar was a constant lunch companion and served as a great role model for me in my PhD work. And if we weren't talking about science we could always debate politics. Thank you to Marc Abrams, Kathy Andrews, Ying Hefner and Helder Balelo who offered essential support along the way. And of course all of the others SBRGers I've been able to work and publish with throughout my time here: Ali Ebrahim, Ramy Aziz, Liz Brunk, Ryan LaCroix, Edward O'Brien, Douglas McCloskey, Daniel Zielinski, Zak King, Gaby Guzman, Donghyuk Kim, Laurence Yang, Juan Nogales, Emanuele Bosi and Miguel Campodonico. Additionally, I'd like to thank all SBRG members of the past, whose work has laid the foundation for mine and who always offered support and insight when I met them at a conference or event. βΩπ for life.

I have to thank all of my friends, back home in Oregon, from the East Coast and here in California for the fun, absurdity and good times we've shared over the years. I'm sorry if sometimes I've dropped off the face of the earth during my time as a PhD student. Hopefully I can make it up to all of you soon. Preferably over a beer.

Finally, I thank my family. My parents have always supported and encouraged me to push my own limits. Their support has been invaluable, both in and out of science. I am lucky to have parents that have pursued areas related to this thesis: data science and medicine. I don't think it's an accident that this thesis focuses on a topic that intersects these two fields. I also thank my younger brother, now an electrical engineer at Intel Corporation, who's been a great companion to grow up with. I thank my grandparents, aunts, uncles and cousins who have all offered support in their own ways over the years. And, of course, our family dogs, Wilson and Penny.

Chapter 1, in part, is a reprint of the material O'Brien EJ*, Monk JM*, Palsson BO: Using Genome-scale Models to Predict Biological Capabilities. *Cell* 2015, 161(5):971-987. The dissertation author was the primary author of this paper. The other authors were Edward J. O'Brien (equal contributor) and Bernhard O. Palsson.

Chapter 2, in part, is a reprint of the material Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BO: Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A* 2013, 110(50):20338-20343. The dissertation author was the primary author.

Chapter 3, in part, is a reprint of the material Bosi E*, Monk JM*, Aziz RK, Fondi M, Nizet V, Palsson BO. Comparative genome-scale modelling of multiple *S. aureus* strains identifies strain-specific pathogenic characteristics and unique metabolic capabilities. *In Preparation.* The disseration author was the primary author (equally contributing with Emanuele Bosi) of this paper.

Chapter 4, in part, is a reprint of the material Monk JM*****, Tsu B*, Buyuktimkin B, Saier MH, Palsson BO. Comparative metabolic network analysis and modelling of four

Leptospira species provides insight into pathogenesis of Leptospirosis. *In Preparation.* The disseration author was the primary author (equally contributing with Brian Tsu) of this paper.

Chapter 5, in part, is a reprint of the material Monk, JM, Koza A, Campodonico A, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM. Quantifying variation within the bacterial species *E. coli* and the impact on host strain selection. *In Preparation.* The disseration author was the primary author of this paper.

Chapter 6, in part, is a reprint of the material Monk JM*, Nogales J*, Palsson BO: Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014, 32(5):447-452. The disseration author was the primary author (equally contributing with Juan Nogales) of this paper.

Chapter 7, in part, is a reprint of the material Monk JM*, Lloyd CJ*, Brunk L, Mih N, King Z, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism for 2016. *In Preparation.* The disseration author was the primary author (equally contributing with Colton Lloyd) of this paper.

Chapter 8, in part, is a reprint of the material Monk JM, Mih N, Brunk E, Aziz RK, Palsson BO. A genome-scale reconstruction of metabolism and associated protein structures in *Staphlyococcus aureus* USA300. *In Preparation.* The disseration author was the primary author of this paper.

Chapter 9, in part, is a reprint of the material Monk JM, Palsson BO: Genetics. Predicting microbial growth. *Science* 2014, 344(6191):1448-1449.The disseration author was the primary author of this paper.

VITA

2008	Bachelor of Science in Engineering, Chemical Engineering, Princeton University, Princeton, NJ
2008-2010	Associate, American Institute of Chemical Engineers. New York, NY
2010-2015	Graduate Student Researcher, University of California, San Diego
2012	Master of Science, Chemical Engineering, University of California, San Diego
2015	Doctor of Philosophy, Chemical Engineering, University of California, San Diego

PUBLICATIONS

1. **Monk JM**, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BO: Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci USA* 2013, 110(50):20338-20343.

2. Bordbar A, **Monk JM**, King ZA, Palsson BO: Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 2014, 15(2):107-120.

3. **Monk JM***, Nogales J*, Palsson BO: Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014, 32(5):447-452. * Authors contributed equally to this work

4. **Monk JM**, Palsson BO: Genetics. Predicting microbial growth. *Science* 2014, 344(6191):1448-1449.

5. Aziz RK, Khaw VL, **Monk JM**, Brunk E, Lewis R, Loh SI, Mishra A, Nagle AA, Satyanarayana C, Dhakshinamoorthy S *et al*: Model-driven discovery of synergistic inhibitors against E. coli and S. enterica serovar Typhimurium targeting a novel synthetic lethal pair, aldA and prpC. *Front Microbiol* 2015, 6:958.

6. Guzman GI, Utrilla J, Nurk S, Brunk E, **Monk JM**, Ebrahim A, Palsson BO, Feist AM: Model-driven discovery of underground metabolic functions in Escherichia coli. *Proc Natl Acad Sci USA* 2015, 112(3):929-934.

7. O'Brien EJ*, **Monk JM***, Palsson BO: Using Genome-scale Models to Predict Biological Capabilities. *Cell* 2015, 161(5):971-987. *Authors contributed equally to this work

8. Yang L, Tan J, O'Brien EJ, **Monk JM**, Kim D, Li HJ, Charusanti P, Ebrahim A, Lloyd CJ, Yurkovich JT *et al*: Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proc Natl Acad Sci USA* 2015, 112(34):10810-10815.

9. Aziz RK, **Monk JM**, Lewis RM, In Loh S, Mishra A, Abhay Nagle A, Satyanarayana C, Dhakshinamoorthy S, Luche M, Kitchen DB et al: Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. Sci Rep 2015, 5:16025.

10. Fouts DE, Matthias MA, Adhikarla H, Adler B, Berg DE, Bulach D, Buschiazzo A, Chang Y, Galloway RL, Haake D, Haft DH, Hartskeerl E, Ko AI, Levett PN, Matsunaga J, Mechaly AE, **Monk JM**, Nascimento ALT, Nelson KE, Palsson BO, Peacock SJ, Picardeau M, Ricaldi JN, Thaipandungpanit J, Wunder EA, Yang XF, Zhang J and Vinetz JE. What Makes a Bacterial Species Pathogenic?: Comparative Genomic Analysis of the Genus Leptospira. *Under review at Plos. Comp. Bio*.

11. Brunk EB, Mih N, **Monk JM**, Zhang Z, O'Brien E, Bliven SE, Chen KE, Chang RL, Bourne PE, Palsson BO. Systems Biology of the Structural Proteome. *Under review at BMC Systems Biology*.

12. Brunk E, George KW, Alonso-Gutierrez J, Thompson M, Baidoo E, Wang G, Petzol CJ, McCloskey D, **Monk JM**, Yang L, O'Brien EJ, Batth TS, Martín HG, Feist A, Adams PD, Keasling JD, Palsson DO, Lee TS. Characterizing Strain Variation in Engineered *E. coli* Using a Multi-omics Based Workflow. *Submitted to Cell Systems.*

13. Bosi E*, **Monk JM***, Aziz RK, Fondi M, Nizet V, Palsson BO. Comparative genome-scale modelling of multiple *S. aureus* strains identifies strain-specific pathogenic characteristics and unique metabolic capabilities. *In Preparation.* *Authors contributed equally to this work.

14. **Monk, JM**, Koza A, Campodonico A, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM. Quantifying variation within the bacterial species *E. coli* and the impact on host strain selection. *In Preparation.*

15. **Monk JM*,** Tsu B*, Buyuktimkin B, Saier MH, Palsson BO. Comparative metabolic network analysis and modelling of four Leptospira species provides insight into pathogenesis of Leptospirosis. *In Preparation.* *Authors contributed equally to this work

16. **Monk JM***, Lloyd CJ*, Brunk L, Mih N, King Z, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism for 2016. *In Preparation.* *Authors contributed equally to this work

17. **Monk JM**, Mih N, Brunk E, Aziz RK, Palsson BO. A genome-scale reconstruction of metabolism and associated protein structures in *Staphlyococcus aureus* USA300. *In Preparation.*

18. Andrews KA; Brunk E; **Monk JM**; Palsson BO; Charusanti P. Genetic basis of growth adaptation after loss of pabC and ubiC in Escherichia coli. *In Preparation*.

19. Aziz KA, **Monk JM**, Andrews K, Nahn J, Khaw V, Wong H, Palsson BO, Charusanti P. The aldehyde dehydrogenase, AldA, is essential for L-1,2-propanediol degradation in laboratory-evolved Escherichia coli. *In Preparation.*

HONORS AND AWARDS

2013 National Science Foundation Fellow, East Asian and Pacific Summer Institute

2015 Outstanding Graduate Student Award, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Comparative Systems Biology Analysis of Microbial Pathogens

by

Jonathan Mayock Monk

Doctor of Philosophy in Chemical Engineering

University of California, San Diego, 2015

Professor Bernhard Ø. Palsson, Chair

Professor Michael Heller, Co-Chair

Genome-scale models of microbial metabolism have become a commonly used tool in the systems biology toolbox. These tools can predict, based solely on an organism's genome sequence, its metabolic capabilities and unique phenotypes in different conditions and under unique perturbations. Furthermore, an array of *in-silico* methods have been developed that can be applied to these models to more deeply characterize an organism, re-engineer it and even to design effective ways to interrupt and kill it. This dissertation discusses the creation and analysis of multiple

genome-scale models of metabolism for different microbial pathogens.

Chapter 1 describes aspects of systems biology that are used throughout this thesis. Topics include the theory and practice of metabolic network reconstruction, genome-scale modelling and flux balance analysis.

Chapter 2 focuses on the reconstruction and analysis of multiple genome-scale metabolic reconstructions of diverse *Escherichia coli* strains. The results highlight strain-specific adaptations to nutritional environments.

Chapter 3 details comparative genome-scale modelling of multiple *S. aureus* strains to identify strain-specific pathogenic characteristics and unique metabolic capabilities that are related to infectious capabilities.

Chapter 4 engages in a comparative metabolic network analysis and modelling of four Leptospira species that provide insight into pathogenesis of Leptospirosis.

Chapter 5 examines seven industrially relevant strains of *E. coli* using transcriptomics and genome-scale models to quantifying variation between the strains that will likely have an impact on host strain selection for metabolic engineering applications.

Chapter 6 conducts an in-depth analysis of existing metabolic network reconstructions and identifies areas where they may need further development.

Chapter 7 details the construction of an updated comprehensive and high-quality genome-scale reconstruction for *Escherichia coli* K-12 MG1655. The model is experimentally validated with gene-knockout studies. Extension to the model are provided, including an application involving production of reactive oxygen species.

Chapter 8 describes the reconstruction of a metabolic network and associated three-dimensional protein structures for *Staphlyococcus aureus* USA300. The model is used to examine basic *S. aureus* biochemistry.

Chapter 9 examines the current state and predicted future of systems biology applications for studying, examining and comparing microbial pathogens.

Chapter 1 Using Genome-Scale Models to Predict Biological Capabilities

1.1 Introduction

Bottom-up approaches to systems biology rely on constructing a mechanistic basis for the biochemical and genetic processes that underlie cellular functions. Genome-scale network reconstructions of metabolism are built from all known metabolic reactions and metabolic genes in a target organism. Networks are constructed based on genome annotation, biochemical characterization, and the published scientific literature on the target organism; the latter is sometimes referred to as the bibliome. DNA sequence assembly provides a useful analogy to the process of network reconstruction (Figure 1A). The genome of an organism is systematically assembled from many short DNA stubs, called reads, using sophisticated computer algorithms. Similarly, the reactome of a cell is assembled, or reconstructed, from all the biochemical reactions known or predicted to be present in the target microorganism. Importantly, network reconstruction includes an explicit genetic basis for each biochemical reaction in the reactome as well as information about the genomic location of the gene. Thus, reconstructed networks, or an assembled reactome, for a target organism represents biochemically, genetically, and genomically structured knowledge bases, or BiGG k-bases. Network reconstructions have different biological scope and coverage. They may describe metabolism, protein-protein interactions, regulation, signaling, and other cellular processes, but they have a unifying aspect: an embedded, standardized biochemical and genetic representation amenable to computational analysis.

A network reconstruction can be converted into a mathematical format and thus lend itself to mathematical analysis and computational treatment. Genome-scale models, called GEMs, have been under development for nearly 15 years and have now reached a high level of sophistication. The first GEM was created for *Haemophilus influenza* and

appeared shortly after this first genome was sequenced (Edwards and Palsson, 1999), and GEMs have now grown to the level where they enable predictive biology (Oberhardt et al., 2009, McCloskey et al., 2013a, Bordbar et al., 2014). Here, we will focus on reconstructions of metabolism and the process of converting them into GEMs to produce computational predictions of biological functions.

The fundamentals of the Constraints-Based Reconstruction and Analysis (COBRA) approach and its uses are also described in this primer, which lays out the constraint-based methodology out at four levels. First, there is a textual description of the methods and their applications. Second, visualization is presented in the form of detailed figures to succinctly convey the key concepts and applications. Third, the figure captions contain more detailed information about the computational approaches illustrated in the figures. Fourth, the primer provides a table of selected detailed resources to enable an in-depth review for the keenly interested reader.

The text is organized into six sections, each one addressing a grand challenge in today's world of "big data" biology:

- **Section 1.2** addresses the collection and organization of disparate data types for an organism of interest and the conversion of this information into a biochemical reaction network reconstruction.
- **Section 1.3** focuses on the conversion of biochemical reconstructions into computational models that can be used to predict metabolic capabilities.
- **Section 1.4** explains the validation of qualitative model predictions and their reconciliation with experimental results to discover new biology.
- **Section 1.5** details advanced genome-scale modeling methods used to make quantitative predictions.

- **Section 1.6** highlights the integration of high-throughput "omics" data with genome-scale models.
- **Section 1.7** examines the future of genome-scale modeling and the prospect of extending these principles to processes beyond metabolism, including transcription and translation.

1.2 A network reconstruction is the systematic assembly of knowledge

A large library of scientific publications exists that describe different model organisms' specific molecular features. Molecular biology's focus on knowing much about a limited number of molecular events changed once annotated genome sequences became available, leading to the emergence of a genome-scale point of view. Now, putting all available knowledge about the molecular processes of a target organism in context and linked to its genome sequence has emerged as a grand challenge. Genome-scale network reconstructions were a response to this challenge.

Network reconstructions organize knowledge into a structured format. The reconstruction process treats individual reactions as the basic elements of a network, somewhat similar to a base pair being the smallest element in an assembled DNA sequence (**Figure 1.1**)**.** To implement the metabolic reconstruction process, a series of questions need to be answered for each of the enzymes in a metabolic network: 1) What are the substrates and products? 2) What are the stoichiometric coefficients for each metabolite that participates in the reaction (or reactions) catalyzed by an enzyme? 3) Are these reactions reversible? 4) In what cellular compartment does the reaction occur? 5) What gene(s) encode for the protein (or protein complex) and what is (are) their genomic location(s)? Genes are linked to the proteins they encode and the reactions they

catalyze using the gene-protein-reaction relationship (GPR). All of this information is assembled from a range of sources including organism specific databases, high-throughput data, and primary literature. Establishing a set of the biochemical reactions that constitute a reaction network in the target organism culminates in a database of chemical equations. Reactions are then organized into pathways, pathways into sectors (such as amino acid synthesis), and ultimately into genome-scale networks, akin to reads becoming a full DNA sequence. This process has been described in the form of a 96-step standard operating procedure (Thiele and Palsson, 2010a).

Today, after many years of hard work by many researchers, there exist collections of genome-scale reconstructions (sometimes called GENREs) for a number of target organisms (Monk et al., 2014a, Oberhardt et al., 2011) and established protocols for reconstruction exist (Thiele and Palsson, 2010a) that can be partially automated (Henry et al., 2010a, Agren et al., 2013a).

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 that, through computations, is able to recapitulate whole cell functions. The grand challenge of disparate data integration into a coherent whole is achieved through the formulation of a GEM. A GEM can then compute cellular states such as an optimal growth state. This process is further explored in the next section. A detailed reading list is available in Table 1 on the network reconstruction process and software tools used to facilitate it.

1.3 Converting a genome-scale reconstruction to a computational model.

Before a reconstruction can be used to compute network properties, a subtle, but crucial step must be taken in which a network reconstruction is mathematically represented. This conversion translates a reconstructed network into a chemically accurate mathematical format that becomes the basis for a genome-scale model (**Figure 1.2A**). 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 (**Figure 1.2B**). These stoichiometries impose systemic constraints on the possible flow patterns (called a flux map, or flux distribution) of metabolites through the network. These concepts are detailed below. Imposition of constraints on network functions fundamentally differentiates the COBRA approach from models described by biophysical equations, which require many difficult-to-measure kinetic parameters. Constraints are mathematically represented as equations that represent balances or as inequalities that impose bounds (**Figure 1.2C**). The matrix of stoichiometries imposes flux balance constraints on the network, ensuring that the total amount of any compound being produced must be equal to the total amount being consumed at steady state. Every reaction can also be given upper and lower bounds, which define the maximum and minimum allowable fluxes through the reactions, that in turn are related to the turnover number of the enzyme and its abundance. Once imposed on a network reconstruction, these balances and bounds define a space of allowable flux distributions in a network; the possible rates at which every metabolite is consumed or produced by every reaction in the network. The flux vector, a mathematical object, is a list of all such flux values for a single point in the space. The flux vector represents a ‘state’ of the network that is directly related to the physiological function that the network produces. Many other constraints such as

substrate uptake rates, secretion rates, and other limits on reaction flux can also be imposed, further restricting the possible state that a reconstructed network can take (Reed, 2012). The computed network states that are consistent with all imposed constraints are thus candidate physiological states of the target organisms under the conditions considered. The study of the properties of this space thus becomes an important subject.

Flux balance analysis (FBA) calculates candidate phenotypes. FBA is the oldest COBRA method. It is a mathematical approach for analyzing the flow of metabolites through a metabolic network (Orth et al., 2010c). This approach relies on an assumption of steady-state growth and mass balance (all mass that enters the system must leave). The constraints discussed above take the form of equalities and inequalities to define a polytope (blue area within the illustration in **Figure 1.2C**) that represents all possible flux states of the network given the constraints imposed. Thus, many network states are possible under the given constraints and multiple solutions exist that satisfy the governing equations. The blue area is therefore often called the 'solution space' to denote a mathematical space that is filled with candidate solutions to the network equations given the governing constraints. FBA uses the stated objective to find the solution(s) that optimize the objective function. The solution is found using linear programming, and, as indicated in **Figure 1.2D**, the optimal solution lies at the edges of the solution space impinging up against governing constrains.

The utility of FBA has been increasingly recognized due to its simplicity and extensibility: it requires only the information on metabolic reaction stoichiometry and mass balances around the metabolites under pseudo-steady state assumption. It computes how the flux map must balance to achieve a particular homeostatic state.

However, FBA has limitations. It balances fluxes, but cannot predict metabolite concentrations. Except in some modified forms, FBA does not account for regulatory effects such as activation of enzymes by protein kinases or regulation of gene expression. More details are found in the caption of Figure 2, and computational resources are summarized below that can be deployed to find the optimal state and to study its characteristics.

Models impose constraints and allow prediction. One of the most basic constraints imposed on genome-scale models of metabolism is that of substrate, or nutrient, availability and its uptake rate (**Figure 1.2E**). Metabolites enter and leave the systems through what are termed “exchange reactions” (i.e., active or passive transport mechanisms). These reactions define the extracellular nutritional environment and are either left ‘open’ (to allow a substrate to enter the system at a specified rate) or ‘closed’ (the substrate can only leave the system). Measurements of the rate of exchange with the environment are relatively easy to perform and they prove to be some of the more important constraints placed on the possible functions of reaction networks internal to the cell. More biological- and data-derived constraints can also be imposed on a model. These advanced constraints are detailed in sections 4, 5 and 6.

1.3.1 What is needed to create a new cell?

The next step in converting a network reconstruction to a model is to define what biological function(s) the network can achieve (**Figure 1.2F**). Mathematically, such a statement takes the form of an ‘objective function.’ For predicting growth, the objective is biomass production, that is, the rate at which the network can convert metabolites into all required biomass constituents such as nucleic acids, proteins, and lipids needed to produce biomass. The objective of biomass production is mathematically represented by

a ‘biomass reaction’ that becomes an extra column of coefficients in the stoichiometric matrix. One can formulate a biomass objective function at an increasing level of detail: Basic, Intermediate, and Advanced (Monk et al., 2014a, Feist and Palsson, 2010a). The biomass reaction is scaled so that the flux through it represents the growth rate (μ) of the target organism.

It is important to note that the biomass objective function is determined from measurements of biomass composition, the uptake and secretion rates from measuring the nutrients in the medium, and the model formulation is based on a network reconstruction that is knowledge-based. Thus, the growth rate optimization problem represents “big data” integrated into a structured format and the hypothesis of a biological objective; grow as fast as possible with the resources available. This is a well-defined optimization problem.

GEMs are input-output “flow models” with an explicit genetic basis. The inner workings of a GEM are readily understood conceptually. In a given environment (i.e., where the nutritional inputs are defined) GEMs can be used to compute network outputs. Flux balance analysis (FBA) can computationally trace a fully balanced path through the reactome from the available nutrients to the prerequisite output metabolite. Such calculations are performed as detailed above with an objective function that describes the removal of the target metabolite from the network. The synthesis of biomass in a cell requires the simultaneous removal of about 60-70 different metabolites. Using FBA, a GEM can also 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. This exercise is one of genome-scale accounting of all molecules flowing through the reactome.

Given its simplicity and utility, FBA has become one of the most widely employed computational techniques for the systems-level analysis of living organisms (Bordbar et al., 2014, Lewis et al., 2012a). It has been successfully applied to a multitude of species for modeling their cellular metabolisms (Oberhardt et al., 2009, Feist and Palsson, 2008a, McCloskey et al., 2013a), and therefore, enabled a variety of applications such as metabolic engineering for the over-production of biochemicals (Yim et al., 2011), (Adkins et al., 2012), identification of anti-microbial drug-targets (Kim et al., 2011), and the elucidation of cell–cell interactions , (Bordbar et al., 2010). Further reading and detailed descriptions of FBA and sources for existing genome-scale models are available in Table 1.

1.4 Validation and reconciliation of qualitative model predictions

Ensuring the consistency and accuracy of all the information available for a target organism is a grand challenge of genome-scale biology. Since model predictions are based on a network reconstruction that represents the totality of what is known about a target organism, such predictions are a critical test of our comprehensive understanding of the metabolism for the target organism. Incorrect model predictions can be used for biological discovery by classifying them and understanding their underlying causes. Performing targeted experiments to understand failed predictions is a proven method for systematic discovery of new biochemical knowledge (Orth and Palsson, 2010b). This section will focus on evaluating qualitative model predictions, their outcomes, underlying causes of incorrect predictions, and how to go about correcting them. Section 4 discusses the same process for quantitative model predictions.

1.4.1 Genetic and environmental parameters

Genome-scale models have many genetic and environmental parameters that can be experimentally varied. Altering the composition of the growth media changes environmental parameters. Alteration of genetic parameters is achieved through genome editing methods. Both environmental and genetic parameters are explicit in GEMs and thus the consequence of both types of perturbations can be computed, predicted, and analyzed. The scale of such predictions has grown steadily since the first genome-scale model of *E. coli* appeared in 2000 (Edwards and Palsson, 2000b).

Genome-scale gene essentiality data are available from specific projects or organism-specific databases. One can systematically remove genes from a reconstruction, and thus the corresponding reactions from the reactome, and repeat the growth computation to predict gene essentiality; i.e., if a growth state cannot be computed without a particular gene, the GEM predicts it to be essential (**Figure 1.3A**). Such growth rate predictions of gene deletion strains have gone from a hundred predictions in the year 2000 (Edwards and Palsson, 2000b), to over 100,000 predictions in 2012 (Yamamoto et al., 2009), and may be heading for over a million predictions in just a few years (Monk and Palsson, 2014).

Both environmental and genetic parameters can be varied when performing FBA. The simplicity of computing growth states (i.e., an output) as a function of media composition (i.e., the nutritional inputs) with the selective removal of genes (**Figure 1.3B**) has led to a number of studies that cross environmental parameters with gene deletions. The explicit relationship between a gene and a reaction makes the deletion of genes and their encoding reactions straightforward. You can readily do this for your target organism, provided that you can construct a library of gene deletion strains.

Improved molecular tools for generating knockout collection libraries (Tn-seq, CRISPR systems, etc.) and improved high-throughput methods for measuring knockout phenotypes have enabled a massive scale-up in the number of phenotypes that can be measured.

1.4.2 Classification of model predictions

Computational predictions of outcomes fall into four categories: true-positives, true-negatives, false-positives and false-negatives. The true-positive and true-negative predictions, where computational predictions and experimental outcomes agree, have generally exceeded 80% to 90% for well-characterized target organisms. Going beyond single gene knockouts to double genes knockouts and more, true negative predictions are particularly significant as they indicate model predictions of true genetic, or epistatic, interactions. In a screen of double-gene yeast knockouts, Szappanos et al. found that models could predict 2.8% of negative genetic interactions (Szappanos et al., 2011). While this indicates poor recall of prediction, of these, 50% were correct, indicating that model predictions are highly precise, but may miss several interactions. These missed predictions represent cases that are currently difficult for functional geneticists to understand. For applications where the goal is to have true predictions, such as for antibiotic design, precision is more important than recall.

FBA based models are highly precise because they are good at predicting impossible states (such as when a gene knockout leads to death). This assumes that the network structure is complete, an assumption that can be a problem when promiscuous enzyme activity arises, leading to a reaction with an encoding gene that is not captured in the model. Models have lower accuracy because FBA assumes that all reactions can happen at maximum rates. Model false positives often occur because an

enzyme is either transcriptionally repressed or does not catalyze the designated reaction at a high enough rate (Table 2, Evaluation of Model Predictions). Predictive failure is perhaps of more interest than success as it represents an opportunity for biological discovery. False negative predictions occur when a GEM predicts the inability to grow in a given environment without the deleted gene, but the experiments show growth. This discrepancy indicates that the reconstructed reactome is incomplete. In contrast, false positive predictions occur when a GEM predicts growth but the experiment results in no growth. This outcome indicates possible errors in the knowledge on which the reactome was based, or that a regulatory process is missing that prevents the use of a gene product factored in the computed solution. An example would be regulation that either represses gene expression or a metabolite-enzyme interaction that inhibits the function of an enzyme that the GEM used to compute the predicted growth state.

Prediction failures can be used to systematically (i.e., algorithmically) generate hypotheses addressing the failures. Such hypotheses have been shown to direct experimentation to improve our knowledge base for the target organism. Computations that vary environmental and genetic parameters become part of a workflow (**Figure 1.3C**). The outcome of the workflow is a set of qualitative model predictions of growth or no growth that are then compared to the experimental outcome of a growth screen. Correct predictions align with experimental results, while incorrect predictions do not. The two are then compared and classified into four categories as shown in **Figure 1.3C**. The failure modes lead to systematic experimentation.

1.4.3 Discovering new metabolic capabilities using model false negatives.

Reconciling such discrepancies between predicted and observed growth states is now a proven approach for biological discovery. A series of algorithms have been

developed that have been shown to compute the most likely reasons for failure of prediction that in turn led to a model-guided experimental inquiry and discovery. Furthermore, high-throughput tools such as phenotypic microarrays and robotic instruments are becoming available to screen cells at high rates. Such discoveries are then incorporated into the reconstruction, leading to its iterative improvement.

The discrepancies between GEM predictions and experimental data have been used to design targeted experiments that correct inaccuracies in metabolic knowledge. In this subsection we provide three illustrative examples that detail how reconciliation of model errors led to the discovery of new metabolic capabilities in three model organisms.

Human: The activity of open reading frame 103 on chromosome 9 (C9orf103) of the human genome was discovered (Rolfsson et al., 2011a) using established gap-filling protocols (Orth and Palsson, 2010b, Reed et al., 2006b). The authors focused on unconnected, “dead end” metabolites in the human metabolic network reconstruction, Recon 1 (Duarte et al., 2007a). Dead end metabolites lead to model errors by creating blocked reactions due to a violation of mass balance. Any flux leading to them cannot leave the network. In an attempt to connect these dead end metabolites, a universal database of metabolic reactions was used to predict the fewest reactions required to fully connect all metabolites in the network. Focusing on gluconate, which is a disconnected metabolite, the authors experimentally characterized (C9orf103), previously identified as a candidate tumor suppressor gene, as the gene that encodes gluconokinase, thereby consuming this metabolite and connecting it to the rest of the human metabolic network.

E. coli: Gap-filling methods combined with systematic gene knockouts in *E. coli* (Nakahigashi et al., 2009b), were used to discover new metabolic functions for the classic glycolytic enzymes phosphofructokinase and aldolase. Single, double, and triple knockout strains of central metabolic genes were grown on 13 different carbon sources. Concurrently, the same gene knockouts and growth conditions were simulated using the *E. coli* GEM. Several discrepancies between model predictions and experimental results were related to *talAB* interactions in the pentose phosphate pathway and could not be reconciled. A metabolomic analysis identified a new metabolite, sedoheptulose-1,7-bisphosphate, that had not been previously characterized. Using metabolic flux analysis and *in vitro* enzyme assays, the investigators confirmed that phosphofructokinase carries out the reaction and that glycolytic aldolase can split the seven-carbon sugar into three- and four-carbon sugars, glyceraldehyde-3-phosphate (G3P) and D-erythrose 4-phosphate (E4P) respectively.

Yeast: An analysis of synthetic lethal screens and gap-filling methods were used to correct incorrect pathways leading to NAD^+ synthesis in yeast (Szappanos et al., 2011). The study compared an experimental set of genetic interactions for metabolic genes against interactions that were predicted by FBA. Using machine-learning techniques, key changes to the metabolic network that improved model accuracy were identified. Model refinement identified one of the two NAD^+ biosynthetic pathways from amino acids in the GEM as a source of inaccurate predictions. Using growth screens with mutant strains, the authors validated that the synthesis of NAD^+ from amino acids was only possible from L-tryptophan (L-trp) but not from L-aspartate (L-asp).

1.4.4 Adaptive laboratory evolution can be used as a part of the discovery process.

In contrast to false negatives, false positives arise when the model predicts growth, but experiments show no growth. False positives occur in cases where experimental data show a particular gene to be essential but model simulations do not **Figure 1.3D**. Metabolic models can be used to predict efficient compensatory pathways, after which cloning and overexpression of these pathways are performed to investigate whether they restore growth and to help determine why these compensatory pathways are not active in mutant cells.

Discovering context-specific regulatory interactions using false positive predictions. Cloning and overexpression of a false positive associated gene has been demonstrated for a *ppc* knockout of *Salmonella enterica* serovar Typhimurium (Fong et al., 2013). A metabolic model of *S.* Typhimurium predicted that the cells could route flux through the glyoxylate shunt when *ppc* is removed due to the backup function of isocitrate lyase encoded by aceA. However, the Δ*ppc* cells were nonviable experimentally. The protein IclR is a transcription factor that regulates the transcription of genes involved in the glyoxylate shunt, including aceA. Therefore a dual knockout ΔppcΔiclR mutant was constructed. Growth was restored in this double mutant at ~60% of the wild type growth rate. Therefore, the prediction of the metabolic model of *S. Typhimurium* failed because it erroneously allowed flux through the glyoxylate shunt when ppc was deleted due to the absence of regulatory information in the model.

Adaptive laboratory evolution can also be used to reconcile false positive predictions. Often, cell populations may need time to adapt to a genetic change or shift in media conditions, giving them the appearance of slow or no growth, despite a model prediction of growth. However, it has been shown that incorrect predictions of *in silico* models based on optimal performance criteria may be incorrect due to incomplete

adaptive laboratory evolution under the conditions examined. It has been shown that *E. coli* K-12 grown on glycerol over 40 days (or about 700 generations) and subjected to a growth rate selection pressure (passing a small fraction of the fastest growers) achieves a final growth rate that is predicted by the GEM (Ibarra et al., 2002). The quantitative prediction of growth rates is discussed in section 4. Thus, a false positive result may indicate that the model is in fact correct and a researcher should be patient while the cell adapts to achieve the model-predicted growth.

Since our knowledge of any target organism is incomplete, its network reconstruction will also be incomplete. Thus, failures in GEM prediction of qualitative outcomes of growth capability are informative about the completeness of a network reconstruction and the consistency of its content. Furthermore, these approaches can be extended beyond model improvement. As genome editing techniques improve, *in silico* prediction of the effect of multiple gene-knockouts will be vital for contextualizing results of knockout studies and engineering genomes to achieve a desired phenotype (Campodonico et al., 2014). Additionally, reconciliation of model false negatives have been used to explore the role that underground metabolism plays in adapting to alternate nutrient environments (Notebaart et al., 2014). The algorithmic procedures that have been developed to address failure of prediction have led to some computer-generated hypotheses resulting in productive experimental undertaking. Further reading about the gap-filling process and algorithms for its implementation are available in Table 1.

1.5 Quantitative phenotype prediction through optimality principles

The previous section treated qualitative predictions that relate to the presence or absence of parts from a reconstruction. Quantitative predictions of phenotypic functions

are more challenging, but possible. The ability to compute quantitative organism functions from a genome-scale model represents a grand challenge in systems biology. Quantitative predictions are achievable with GEMs (even if they are based on incomplete reconstructions) by deploying cellular optimality principles. Evolutionary arguments underlie the deployment of optimality-based hypotheses. Phenotypes maximizing a hypothesized fitness function (as represented by an objective function) can be computed with constrained-optimization methods (Orth et al., 2010c).

As for qualitative binary predictions of possible growth states, incorrect quantitative predictions often lead to new biological hypotheses and understanding. However, the discoveries arising from quantitative phenotype predictions are typically of a different nature than qualitative predictions. Rather than relating to missing reconstruction content (Section 3), the discoveries from quantitative phenotype prediction often relate to broad, fundamental organismal constraints (Beg et al., 2007, Zhuang et al., 2011b) and evolutionary objectives and trade-offs (Shoval et al., 2012).

Quantitative phenotype prediction has also proven to be a useful capability for bioengineering applications. By optimizing an engineering (instead of evolutionary) objective, the best possible performance of an engineered biological system can be determined. Furthermore, the specific flux states needed to achieve high performance can guide engineering design (King et al., 2015b).

1.5.1 Workflow for quantitative phenotype prediction.

Quantitative phenotypes can be predicted through the same computational procedures used for qualitative growth predictions (**Figure 1.4A**). An objective (either evolutionary or engineering) is assumed, and maximized computationally (subject to flux balance and other constraints). The flux state(s) that maximize the objective are then the

predicted quantitative fluxes. These predictions can then be compared to experimental measurements. In cases of agreement, the evolutionary hypothesis is supported. In cases of a disagreement between experimental and theoretical predictions, either: a) the biological system has not been exposed to the selection pressure to reach the theoretical optimum (i.e., the assumed evolutionary objective is incorrect or partially correct), or b) there are missing biological constraints that affect the theoretical predictions (i.e., the relevant biological constraints are incomplete).

Experimental evolution can discriminate these alternatives (Ibarra et al., 2002, Schuetz et al., 2012) by exposing the biological system to the appropriate selection pressure, leading it to evolve towards the stated optimum. For example, in one study, strains carrying deletions of one of six metabolic genes were evolved on four different carbon sources. A total of 78% of strains tested reached the metabolic model predicted optimal growth rate after adaptive laboratory evolution after 40 days of passage (Fong and Palsson, 2004).

1.5.2 Flux variability analysis (FVA) calculates possible flux states.

Flux balance analysis computes an optimal objective value and a flux state that is consistent with that objective (and all of the imposed constraints). While the objective value is unique, multiple flux states can typically support the same objective value in genome-scale models. For this reason, flux variability analysis (FVA) is used to determine the possible ranges for each reaction flux (Mahadevan and Schilling, 2003b). With FVA, the objective value is set to be equal to its maximum value, and each reaction is maximized and minimized. For some fluxes, their maximum value will be equal to their minimum, enabling a specific prediction. For others, there may be a wide range of possible values due to alternative pathways. Often, a parsimonious flux state is also

assumed and computed with parsimonious-FBA (pFBA) (Lewis et al., 2010a). With pFBA, the sum of fluxes across the entire network is minimized (again, subject to the optimal objective value determined); pFBA will eliminate some alternative pathways. Typically, many reaction fluxes can by uniquely predicted with optimality and parsimony assumptions. Additional biological constraints in next-generation models (Section 6) reduce the possible flux states further (Lerman et al., 2012a).

1.5.3 Types of possible (evolutionarily optimal) quantitative predictions.

The simplest type of quantitative phenotype predictable with constraint-based models is nutrient utilization. While metabolic models do not predict absolute rates of nutrient uptake, they predict the optimal ratios at which nutrients are utilized. For example, metabolic models predict an optimal oxygen uptake rate relative to the carbon source uptake rate (resulting in a predicted optimal ratio between the two nutrients). In an early study, the ratios of oxygen and carbon uptake were shown to be predictable for a number of carbon sources in *E. coli* (Edwards et al., 2001). In a later study, *E. coli* was evolved in the laboratory on a carbon source (glycerol) for which the wild-type strain did not match the predicted nutrient utilization; after evolution, the strain exhibited the optimal uptake rates predicted theoretically (**Figure 1.4B**) (Ibarra et al., 2002). Comparison of experimental and predicted phenotypes therefore reveals the environments to which an organism has been evolutionary exposed.

Metabolic fluxes for central carbon metabolism can be estimated with ^{13}C carbon labeling experiments, making them candidates for quantitative prediction (**Figure 1.4B**). Since the dimensionality of carbon labeling data is larger than that for nutrient uptake, there is more opportunity to dissect the differences in computed and measured fluxes to better understand the multiple objectives and constraints underlying microbial

metabolism. Impressively, the biomass objective function can explain a large amount of the variability of fluxes (Schuetz et al., 2007). Failure modes in prediction have led to the appreciation of the importance of protein cost (O'Brien et al., 2013a), and membrane (Zhuang et al., 2011b) and cytoplasmic spatial constraints (Beg et al., 2007), which affect the optimal flux state (**Figure 1.4C**). Furthermore, failure modes have led to the understanding that metabolism is simultaneously subject to multiple competing evolutionary objectives, resulting in trade-offs (e.g., growth versus maintenance) employed by different species (**Figure 1.4C**). In this way, outliers in quantitative predictions can improve the understanding of constraints and objectives underlying a particular organism's metabolism.

Optimality principles from stoichiometric models have also been expanded from single populations of cells to microbial communities. To model microbial communities, multiple species are linked together through the exchange of nutrients extra-cellularly (Stolyar et al., 2007) or through direct electron transfer (Nagarajan et al., 2013). The secretion rate from one species limits the uptake rate for others, resulting in balanced species interactions. For a number of cases of communities composed of two or three members, the optimal rate of nutrient exchange and the ratio of the species in the population (Wintermute and Silver, 2010) can be predicted. The effects of spatial organization of community members are also being uncovered (Harcombe et al., 2014). The constraints on nutrient flow between organisms (e.g., diffusion) have proven to be important for predicting community composition and behavior, highlighting the importance of abiotic constraints and community structure in the behavior of biological communities.

Evolution is a natural counterpart to optimality-based predictions with constraint-based methods. Constraint-based optimality predictions have focused on predicting the endpoints of short-term experimental evolution. However, this scope of application has increased in recent years to study long-term phenotypic and enzyme evolution (Nam et al., 2012, Plata et al., 2015).

1.5.4 From optimality principles to prospective design

Quantitative phenotype prediction via optimization is also commonly used for bioengineering applications (**Figure 1.4D**). For example, in metabolic engineering, optimal pathway yields are used to prioritize pathways to be built into a production strain and to benchmark their performance. Furthermore, the flux states required to achieve these optima (and how they differ from wild-type growth states) can guide strain design (Cvijovic et al., 2011).

A number of design algorithms have been built to work with metabolic models and predict the genetic and environmental modifications to increase performance (Ranganathan et al., 2010, Burgard et al., 2003). While many design algorithms and applications have been focused on metabolite production (e.g., for production of fuels and chemicals), metabolic models have also been utilized for the design of biosensors (Tepper and Shlomi, 2011) and biodegradation (Scheibe et al., 2009, Zhuang et al., 2011a). Also, design has expanded beyond single populations to microbial communities/ecosystems (Klitgord and Segre, 2010).

Quantitative phenotype predictions initially focused on simple physiological predictions and are still expanding to more complex phenotypes, biological systems (Levy and Borenstein, 2013), and environments. Although there have been notable successes of quantitative phenotype prediction, certain phenotypes are still difficult to

predict. Historically, difficult predictions have led to the development of new computational methods and an appreciation of new biological constraints. Table 2 (Evaluation of Model Capabilities) summarizes several types of predictions and the approximate performance of constraint-based methods utilized to date. The expansion in the scope and accuracy of predictions continues today with models of increased scope (O'Brien et al., 2013a, Chang et al., 2013a), discussed in section 6.

Thus far, quantitative phenotypes have been limited primarily to microbial systems and, more recently, plants (Williams et al., 2010, Collakova et al., 2012). For multi-cellular organisms, specialized cell types support the fitness of the entire organism. Cell-type specific 'objectives' have been constructed (Chang et al., 2010), though they typically are used for qualitative (Section 3) instead of quantitative phenotype prediction. Instead, quantitative phenotypes in multi-cellular organisms are typically determined through model-driven analysis of experimental data, discussed in Section 5.

1.6 Multi-omic data integration: constraining and exploring possible phenotypic states

With the expanding quantity of omics and other phenotypic data, there is an increasing need to integrate these datasets to drive further understanding and hypothesis generation. Phenotypic data types can be integrated with metabolic GEMs to determine condition-specific capabilities and flux states in the absence of assumed objectives (Section 4). Computational methods that identify the possible range of phenotypic states given the measured data allow one to quantify the degree of (un)certainty in metabolic fluxes. Some types of data are quantitative and directly indicative of metabolic fluxes, whereas other data are qualitative or indirectly related to metabolic fluxes. By layering different data types, the true state of a biological system

can be determined with increased precision. The need for formal integration of disparate data types represents a grand challenge that has been termed Big Data to Knowledge (BD2K, bd2k.nih.gov).

1.6.1 Workflow for multi-omic data integration.

The overall procedure for multi-omic integration with genome-scale models is an iterative workflow (**Figure 1.5A**). Once experimental data from the particular biological system under study is obtained, it is converted into constraints on model function (**Figure 1.5B**). The successive application of experimentally derived constraints to the reaction network results in the generation of a cell-type and condition-specific model (**Figure 1.5C**). Several computational procedures can then be used to explore the metabolic capabilities and achievable phenotypes of the experimentally constrained model (**Figure 1.5D**). Evaluation of these phenotypic capabilities and comparison of different cells or environments leads to identification of their molecular differences (**Figure 1.5E**), providing biological insight and driving further hypotheses.

1.6.2 Converting data to model constraints.

Successive imposition of constraints is a basic principle of COBRA (Palsson, 2000). Some data types can be directly converted into constraints on model variables. Biomass composition and growth rate affect the metabolic demands of cellular growth (Feist and Palsson, 2010a). Time-course exo-metabolomics can be used to set the uptake and secretion rates of nutrients (Mo et al., 2009). Intracellular quantitative metabolomics combined with reaction free energies can discern condition-specific reaction directionalities (Henry et al., 2007). Isotopomer distributions from cellular biomass or metabolite pools can be used to infer and constrain intracellular fluxes

(Zamboni et al., 2009). These data can be used separately or combined to identify with increasing precision the true state of the cell.

Other data types affect metabolism more qualitatively. In theory, quantitative metabolite, transcript, and protein levels can be used to constrain metabolism quantitatively, but in practice this requires many parameters that are hard-to-measure and are organism-specific. Instead, these data types can be used as qualitative constraints relating to gene product or metabolite presence/absence; that is, if a metabolite is present, a reaction must be active that produces it (Shlomi et al., 2008), and if a gene product is absent, it's catalyzed reactions cannot carry flux (Jerby et al., 2010, Schmidt et al., 2013a). Similarly, regulatory interactions can be added to affect the presence/absence of a gene product based on condition-specific activity of a transcription factor (Chandrasekaran and Price, 2010).

1.6.3 Cell-type and condition-specific models

Starting from a large reconstructed reaction network (e.g., representing all metabolic reactions encoded in the human genome (Thiele et al., 2013b)), the imposition of experimental data results in the generation of cell-type and condition-specific models. Experimentally derived constraints pare down the achievable phenotypes from those encoded by the totality of the cell's genome. By eliminating phenotypes that cannot be achieved, this new model represents the capabilities of the particular cell-type and environment assayed. This model summarizes the experimental data in a self-consistent and integrated format, and forms the starting point for further computational and biological inquiry (Shlomi et al., 2008, Agren et al., 2012) (see **Figure 1.5D,E**).

1.6.4 Quantifying uncertainty with Flux variability analysis (FVA) and Sampling.

Once a cell-type and condition-specific model is created, computational methods are used to determine the possible flux states of the cell. FVA (which is described in section 4) (Mahadevan and Schilling, 2003b) can be used to determine the range of fluxes that are consistent with the experimental data. A more refined approach is flux sampling (Schellenberger and Palsson, 2009a) (typically with Markov Chain Monte Carlo, MCMC, methods), which determines the distribution of fluxes for all reactions (instead of simply the range). When no cellular objective is assumed, the feasible flux space is very unconstrained and a particular reaction could be operating at nearly any flux value. As more data is layered, the feasible flux space decreases. When no objective is assumed, fluxes are rarely precisely known, and many will remain completely unknown. However, an imprecisely known flux space is often sufficient to discern differences between two environments/states as discussed in the following subsection.

1.6.5 Using computed states to drive discovery and experimentation.

Once the range of possible phenotypic states is quantified, they must be analyzed to gain biological insights. Often a comparative approach is employed, in which two experimental states (e.g., neurons from Alzheimer's disease patients compared to healthy controls (Lewis et al., 2010b)) are compared. Reactions that have a non-overlapping FVA range must be different between the two states, and can be indicative of important metabolic changes. In cases where the FVA ranges are overlapping, the flux distributions from MCMC sampling can still be different – that is, the reactions are likely different between the two states, but the current experimental data is insufficient to guarantee it.

Pathway visualization is also helpful in gaining insight into changes in cell states—fluxes (or flux ranges) are most comprehensible in a network context. A few tools exist for the visualization of metabolic fluxes; some are based on static maps (Schellenberger et al., 2010), whereas others create auto-generated layouts and new tools allow for the drawing of maps based on flux solutions (King and Ebrahim, 2014). Finally, identifying reactions or subsystems that remain partially identified (e.g., based on a large FVA range) can guide further experimentation, resulting in an iterative computational and experimental elucidation of a cell's state.

GEMs can be used to integrate numerous data types. In fact, as more experimentally derived constraints are successively imposed, analysis often becomes easier (as the range of possible solutions shrinks (Reed, 2012)), instead of more challenging as often occurs with statistically based data integration procedures. A current challenge with metabolic GEMs is the explicit integration of data types that do not directly reflect metabolic fluxes (e.g., transcriptomics, proteomics, and regulatory interactions). This challenge is primarily due to the fact that these processes are not explicitly described in metabolic models. Expansions of metabolic models to encompass gene expression hold promise to address this challenge and are discussed in section 6.

1.7 Moving beyond metabolism to molecular biology

Up to this point, this primer has focused on metabolic models, or M-Models. M-models have reached a high degree of sophistication after 15 years of development, resulting in standard operating procedures for their construction (Thiele and Palsson, 2010a) and use (Schellenberger et al., 2011b). However, M-Models are limited in their explicit coverage to metabolic fluxes. Thus, a grand challenge in the field has been to expand the concepts of constraint-based models of metabolism to other cellular

processes to formally include more disparate data types in genome-scale models (Reed and Palsson, 2003a).

1.7.1 Computing properties of the proteome.

The process of addressing this grand challenge has begun. Recently, genome-scale network reconstructions have expanded to encompass aspects of molecular biology (**Figure 1.6A**). Two significant expansions are genome-scale models integrated with protein structures, GEM-PRO, and integrated models of metabolism and protein expression, ME-Models. GEM-PRO allows for structural bioinformatics analysis to be performed from a systems-level perspective, and have those results in turn affect network simulations. ME-Models allow for the simulation of proteome synthesis, and account for the capacity and metabolic requirements of gene expression.

1.7.2 GEM-PRO -- A structural biology view of cellular networks

GEM-PRO reconstructions can have varying degrees of detail, which affects the types of analysis possible (**Figure 1.6B**). So far, GEM-PRO reconstructions have been created for *T. maritima* (Zhang et al., 2009) and *E. coli* (Chang et al., 2013a, Chang et al., 2013b). Initial reconstructions have focused on single peptide chains (Zhang et al., 2009), and utilized homology modeling to fill in gaps where organism-specific structures have not been identified. Further reconstruction detail has included protein-ligand complexes (Chang et al., 2013a) and quaternary protein assemblies (Chang et al., 2013b). To link the structures to the metabolic model, structural data directly references the GPRs in the metabolic reconstruction. For cases of protein-metabolite complexes, the metabolites also need to be properly annotated in the structural data. The structural reconstruction therefore provides a physical embodiment of the gene-protein-reaction relationship.

There are a few notable cases demonstrating the unique analysis possible with the combination of protein structures and network models. In *T. maritima*, network context and protein fold annotations were combined to test alternative models for pathway evolution (Zhang et al., 2009). The *T. maritima* GEM-PRO supported the patchwork model for genesis of new metabolic pathways. In *E. coli*, the effect of temperature on protein stability and enzyme activity was simulated at the systems level, recapitulating the effects of temperature on growth (Chang et al., 2013a). Also in *E. coli*, protein-ligand interactions were combined with gene essentiality predictions to discover new antibiotic leads and off-targets (Chang et al., 2013b). These examples just scratch the surface of analyses made possible with the integration of network and structural biology.

1.7.3 Modeling molecular biology and metabolism with ME-Models.

ME-Models formalize all of the requirements for biosynthesis of the functional proteome. They compute the proteome composition and its integrated function to produce phenotypic states and all the metabolic processes needed for its synthesis. This represents an integrated view of metabolic biochemistry and the core processes of molecular biology. As with GEM-PRO, the first ME-Models were formulated for *T. maritima* (Lerman et al., 2012a) and *E. coli* (O'Brien et al., 2013a, Thiele et al., 2012).

The reconstruction of a ME-Model starts with the formation of reactions for gene expression and enzyme synthesis (Thiele et al., 2009b). The processes explicitly accounted for in ME-Models are very detailed, including transcription units and initiation and termination factors for transcription, tRNAs and chaperones needed for translation and protein folding, and metal ion and prosthetic group requirements for catalysis. In other words, the reconstructions strive to match as closely as possible all the

biochemical processes required to synthesize fully functional enzymes. To create a ME-Model, the reactions for enzyme synthesis are coupled to the totality of metabolic reactions with pseudo-kinetic constraints, termed ‘coupling constraints’ (Thiele et al., 2010, Lerman et al., 2012a). These constraints relate the abundance of an enzyme (or any ‘recyclable’ chemical species, e.g., mRNA, tRNA), to its degradation rate and catalytic capacity.

ME-Models thus significantly expand the scope of phenotype predictions possible to include aspects of transcription and translation. RNA and protein biomass composition are variables in ME-Models, and are no longer set *a priori* (as in the biomass objective function of M-Models). ME-Models predict the experimentally observed linear changes in the ratio of RNA-to-protein mass fractions as a consequence of changes in protein synthesis demands (O'Brien et al., 2013a). Furthermore, the mass fractions of protein subsystems agree well with those predicted by the ME-Model. This shows that the broad distribution of protein subsystem abundance is predictable using optimality principles and the comparison reveals that some subsystems were under-predicted, thus identifying them as gaps in knowledge and targets for further reconstruction and model refinement (Liu et al., 2014b). While the quantitative prediction of individual protein abundances is currently out of scope of the ME-Model (as these demands depend on enzyme-specific kinetics) the ME-Model has been shown to accurately predict differential expression across certain environmental shifts, due to the differential requirements of proteins across conditions (a more qualitative than quantitative prediction) (Lerman et al., 2012a).

A recent expansion to the ME-Model includes the addition of protein translocation, allowing for the localization of protein to be computed (Liu et al., 2014b)

(i.e., into cytoplasm, periplasm, inner and outer membrane). Translocase abundances and compartmentalized proteome mass was accurately predicted from the bottom-up based on optimality principles. Addition of compartmentalization also allows for membrane area and cytoplasmic volume constraints to be formalized, which, if combined with GEM-PRO, approaches a digital embodiment of a three-dimensional cell.

Metabolic models are limited in their predictive ability dictated by the scope of the reconstruction. Nearly all of the predictions of metabolic models outlined in the previous sections can be refined and expanded with GEM-PRO or ME-Models. Advances to include protein structures and protein synthesis open new vistas for constraint-based modeling.

The scope of genetic perturbations (Section 2) that can be simulated is significantly larger due to the inclusion of genes for gene expression (and accounting for protein cost) and the effects of coding mutations on protein structures; GEM-PRO also expands the scope of environmental perturbation to enable simulation of changes in temperature. GEM-PRO allows for new gap-filling approaches (Section 3) based on structural bioinformatics methods. ME-Models expand the scope of quantitative molecular phenotypes to include transcript and protein levels (Section 4), and transcriptomics and proteomics can be analyzed in mechanistic detail (Section 5).

With the added capabilities of GEM-PRO and ME-Models also come additional computational challenges. While single optimization calculations with M-Models take less than a second on a modest laptop computer, growth-maximization with a ME-Model can take over an hour. The ME-Model also requires specialized high-precision solvers. Many promising applications of GEM-PRO will require simulation of protein dynamics with molecular dynamics (MD) and hybrid quantum mechanics/molecular mechanics

(QM/MM) simulations on protein structures. High-performance computing environments are required for such simulations, and there is a pervasive trade-off between the precision of simulations and the scope of structural coverage. However, advances in high-precision solvers for ME-Models (Sun et al., 2013) and structural simulations for GEM-PRO are rapid and are likely to ameliorate these challenges.

Like discoveries enabled by comparing M-Model predictions to experimental data, we anticipate much biology can be learned from comparing *in silico* and *in vivo* proteome allocation (O'Brien and Palsson, 2015), leading to increasingly predictive models. The *E. coli* ME-Model currently encompasses many key cellular functions, covering ~80% of the proteome by mass in conditions of exponential growth; the remaining proteome mass outside of the scope of the model can guide model expansion. In addition to DNA replication and cell division (Karr et al., 2012), much of the remaining proteome mass involves cellular stress responses (e.g., pH, osmolarity, osmotic); like with temperature, GEM-PRO will aid in modeling these cellular stresses.

1.8 Perspective

Genome-scale models have been under development since the first annotated genome-sequences appeared in the mid to late 1990s. For most of this history, the focus of GEMs has been on metabolism. After initial successes with metabolic GEMs it became clear that the same approach could be applied to other cellular process that could be reconstructed in biochemically accurate details. Thus, a vision was laid out in 2003 that the path to whole cell models was conceptually possible and that such models could be used as a context for mechanistically integrating disparate omic data types (Reed and Palsson, 2003a). This vision is now being realized. This primer shows how six grand challenges in cell and molecular and systems biology can be addressed using

GEMs. A surprising range of cellular functions and phenotypic states can be now dealt with.

We now have the tools at hand to develop quantitative genotype-phenotype relationships from first principles and at the genome-scale. Current models of prokaryotes account for metabolism, transcription, translation, protein localization, and protein structure. Processes not described in the current ME models will be systematically reconstructed over the coming years to gain a more and more comprehensive description of cellular functions. Biology can thus look forward to the continued development and use of a mechanistic framework for the study of biological phenomena as physics and chemistry have enjoyed for over a century.

1.9 Acknowledgments

This work was supported by National Institutes of Health grant no. R01 GM057089.

Chapter 1, in part, is a reprint of the material O'Brien EJ*, Monk JM*, Palsson BO: Using Genome-scale Models to Predict Biological Capabilities. *Cell* 2015, 161(5):971-987. The dissertation author was the primary author of this paper. The other authors were Edward J. O'Brien (equal contributor) and Bernhard O. Palsson.

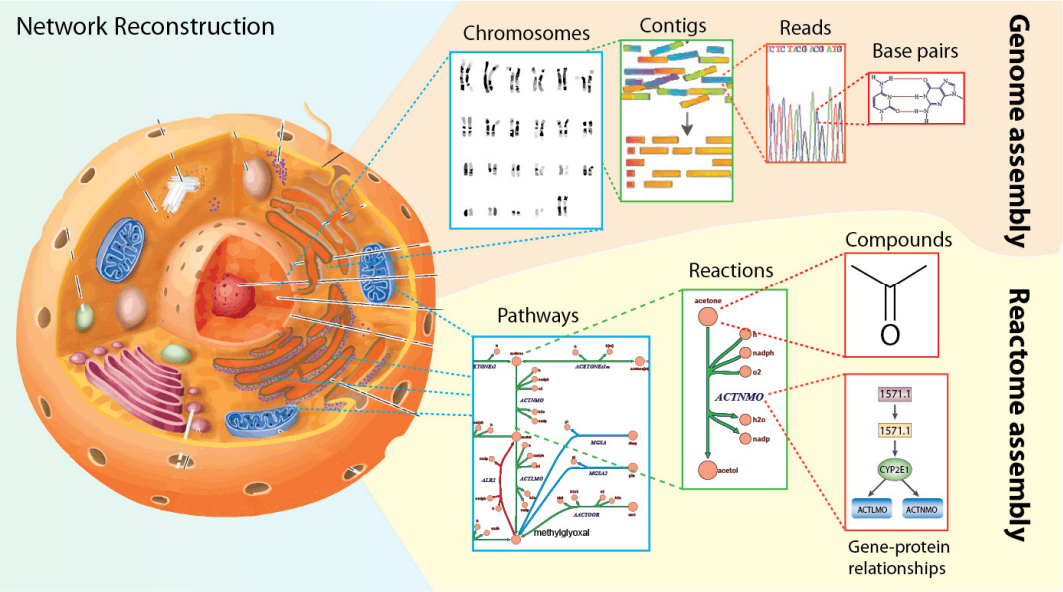


Figure 1.1. Network Reconstruction.
An organism's reactome can be assembled in a way that is analogous to DNA sequencing assembly. From right to left: first the interacting compounds must be identified. Then, the reactions acting on these compounds are tabulated and the protein that catalyzes the reaction and the corresponding open reading frame is identified in the organism of interest. These reactions are assembled into pathways that can be laid out graphically to visualize a cell's metabolic map at the genome-scale. Several tools for reactome assembly and curation exist including the COBRA Toolbox (Schellenberger et al., 2011c, Ebrahim et al., 2013), KEGG (Kanehisa et al., 2014), EcoCyc (Keseler et al., 2013), ModelSeed (Henry et al., 2010c), BiGG (Schellenberger et al., 2010), Rbionet (Thorleifsson and Thiele, 2011), Subliminal (Swainston et al., 2011), Raven toolbox (Agren et al., 2013a) and others.

Figure 1.2. Formulation of a computational model
A) After the metabolic network has been assembled it must be converted into a mathematical representation. This conversion is performed using a stoichiometric (**S**) matrix where the stoichiometry of each metabolite involved in a reaction is enumerated. Reactions form the columns of this matrix and metabolites the rows. Each metabolite's entry corresponds to its stoichiometric coefficient in the corresponding reaction. Negative coefficient substrates are consumed (reactants), and positive coefficients are produced (products). Converting a metabolic network reconstruction to a mathematical formulation can be achieved with several of the toolboxes listed in Table 1. Care must be taken to ensure that all reactions are mass and charge balanced and that all are thermodynamically feasible in the condition of interest, and that their cellular location is accurate. Such checks represent quality controls on the reconstruction to ensure that it is chemically accurate and that the computations from the corresponding GEM are physico-chemically meaningful. B) Constraints can be added to the model such as 1) enforcement of mass balance and 2) reaction flux (v) bounds. The blue polytope represents different possible fluxes for reactions 5 and 6 consistent with stated constraints. Those outside the polytope violate the imposed constraints and are thus 'infeasible.' **C)** At their base, constraint-based models predict the flow of metabolites through a defined network. The predicted path is determined using linear programming solvers and termed Flux Balance Analysis (FBA). FBA can be used to calculate the optimal flow of metabolites from a network input to a network output. The desired output is described by an objective function. If the objective is to optimize flux through reaction 5, the optimal flux distribution would correspond to the levels of flux 5 and flux 6 at the blue point circled in the figure. Alternatively if the objective was to optimize flux through reaction 6, the optimal flux distribution would correspond to the levels of flux 5 and flux 6 at the red point. The objective function can be a simple value or draw on a combination of outputs, such as the biomass objective shown in **Figure 1.2E**. The optimize function in the COBRA toolbox returns the output of the objective function as well as the flux of each reaction in the network contributing to achieve that output. Therefore, the optimizations displayed in the figure would be performed in the COBRA toolbox using flux 5 or 6 as the objective. It is important to note that alternate optimal flux distributions may exist to reach the optimal state as discussed in **Figure 1.4C**. **D)** Once a network reconstruction is converted to a mathematical format, the inputs to the system must be defined by adding consideration of the extracellular environment. Compounds enter and exit the extracellular environment via 'exchange' reactions. The GEM will not be able to import compounds unless a transport reaction from the external environment to the inside of the cell is present. **E)** In addition to exchange reactions, the biomass objective function acts as a drain on cellular components in the same ratios as they are experimentally measured in the biomass. In FBA simulations the biomass function is used to simulate cellular growth. This simulation can be performed using the optimize function in the COBRA toolbox. The biomass function is composed of all necessary compounds needed to create a new cell including DNA, amino acids, lipids and polysaccharides. The biomass objective function is not the only physiological objective that can be examined using COBRA tools.

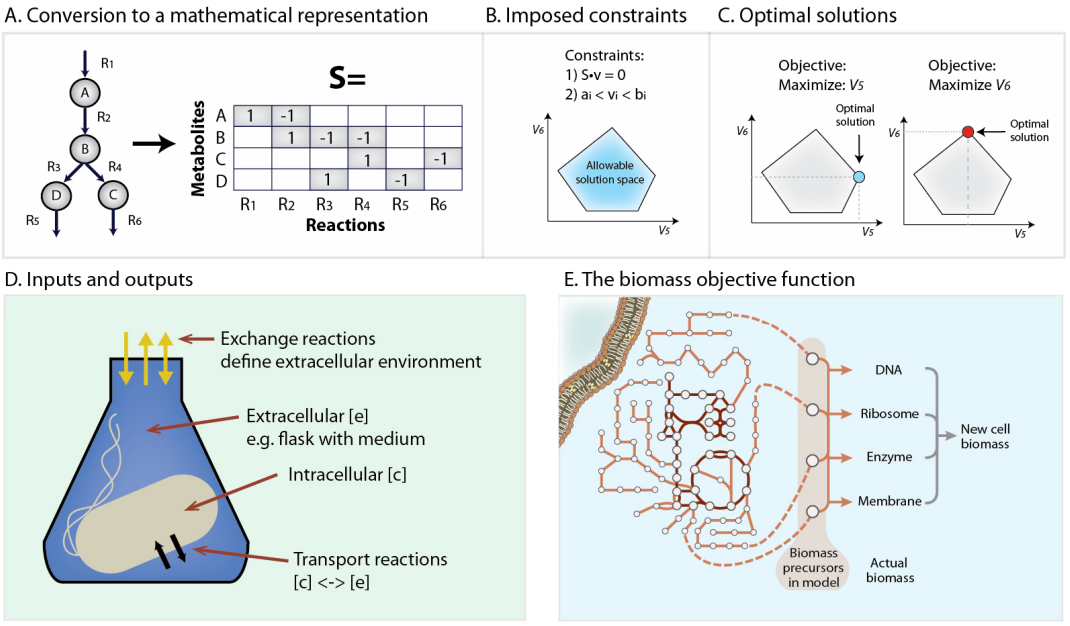
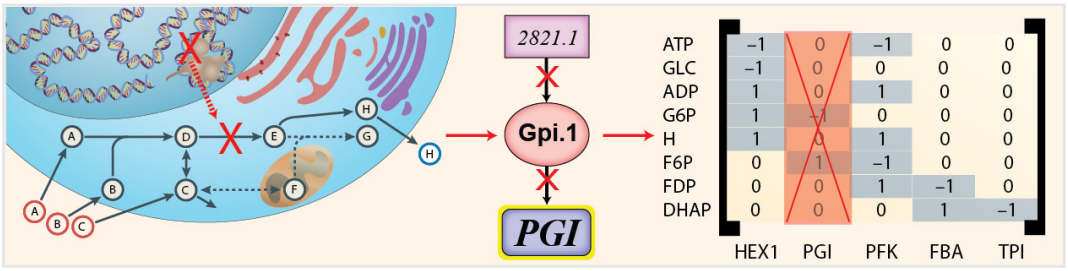
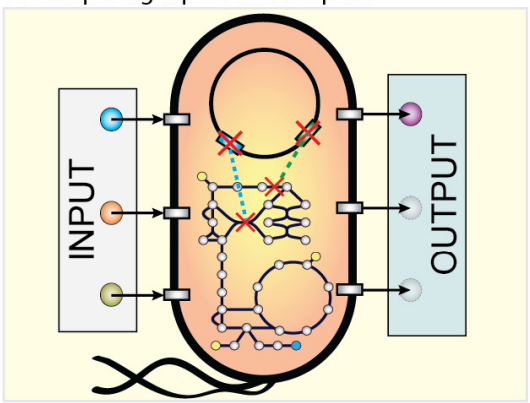

A. Conversion to a mathematical representation
B. Imposed constraints
C. Optimal solutions
R1
A
R2
B
R3
R4
D
C
R5
R6
S=
Metabolites
A
B
C
D
1 -1
1 -1 -1
1 -1
1 -1
R1 R2 R3 R4 R5 R6
Reactions
Constraints:
1) S•v = 0
2) ai < vi < bi
v6
Allowable solution space
v5
Objective:
Maximize: v5
Optimal solution
Objective:
Maximize v6
Optimal solution
D. Inputs and outputs
Exchange reactions
define extracellular environment
Extracellular [e]
e.g. flask with medium
Intracellular [c]
Transport reactions
[c] <-> [e]
E. The biomass objective function
DNA
Ribosome
Enzyme
Membrane
New cell biomass
Biomass precursors in model
Actual biomass

Figure 1.3. Using models for qualitative predictions and iterative improvement
A) Each reaction in the network is linked to a protein and encoding gene through the gene-protein-reaction (GPR) relationship. Because each reaction in the network corresponds to a column in the stoichiometric matrix, simply removing the column association with a particular reaction can simulate gene knockouts. The delete model gene function in the COBRA toolbox can be used to perform these functions automatically and takes into account isozymes and protein complexes. Multiple KO simulations can be performed. An example could easily delete every pairwise combination of 136 central carbon metabolic *E. coli* genes to find double gene knockouts that are essential for survival of the bacteria. **B)** The simplicity of altering inputs to change cellular growth environments (for example using the function changeRxnBound) and removing genes *in silico* allows one to perform simulations in millions of experimental conditions (multiple gene knockouts can be crossed with multiple environments) quickly. Even on a modest laptop computer a single FBA calculation runs in a fraction of a second, thus simulating the effect of all gene knockouts in *E. coli* central metabolism can be run in less than 10 seconds. Such simulations can then be compared to experimental data to classify prediction success and failure modes. **C)** Incorrect model predictions are an opportunity for biological discovery because they highlight where knowledge is missing. Targeted experiments can be performed to discover new content that can then be added back to the model to improve its predictive accuracy. Missing model content can be discovered using automated approaches known as 'gap-filling' (Orth and Palsson, 2010a). The *growMatch* function in the COBRA toolbox can be used to query a database of potential reactions to fill gaps in a network and restore *in silico* growth to a model. **D)** Gap-filling approaches have been used previously to discover new metabolic reactions in several organisms of interest, clockwise from upper left are examples from three model organisms: *E. coli*, human and yeast. *E. coli*: Two new functions for two classical glycolytic enzymes phosphofructokinase (PFK) and fructose-bisphosphate aldolase (FBA) were discovered (red) (Nakahigashi et al., 2009a). Also the *E. coli talA* and *talB* genes were newly discovered to catalyze a transaldolase reaction TALA (blue) based on knockout phenotypes. Human: Gluconokinase (EC 2.7.1.12) activity was discovered based on the known presence of the metabolite 6-phosphogluconolactonate in the human reconstruction(Rolfsson et al., 2011b) (red). Yeast: Automated model refinement suggested modifications in NAD biosynthesis pathway. Experimental results indicated negative genetic interactions in the NAD biosynthesis pathway (Δbna vs WT) starting from tryptophan indicating that a parallel pathway from aspartate thought to exist in yeast was not present (Szappanos et al., 2011). **E)** False positive predictions can be reconciled by adding regulatory rules derived from high throughput data (Covert et al., 2004), for example, a recent study was able to reconcile 2,442 false model predictions from the *E. coli* GEM by updating the function of just 12 genes using the GeneForce algorithm (Barua et al., 2010). Additionally, a false positive growth inconsistency in the metabolic model of *S. Typhimurium* was reconciled by updating regulatory rules for the *iclR* gene product's transcriptional repression of *aceA* encoding isocitrate lyase. Also, transcriptional repression can often be relieved via adaptive laboratory evolution. Such evolution drives experimental phenotypes to achieve model predictions. Several experimental studies have shown that an organism can evolve to eventually achieve the model-predicted optimal growth state (Ibarra et al., 2002).

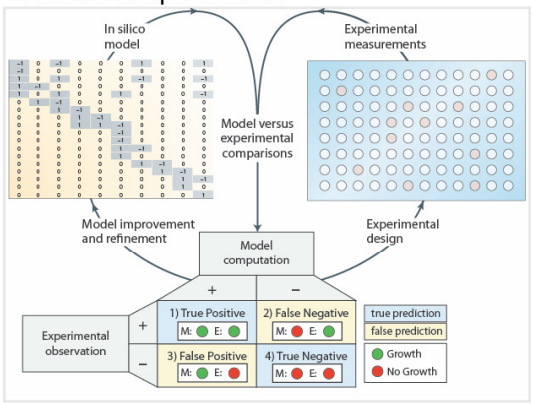
A. Simulating gene and reaction knockouts



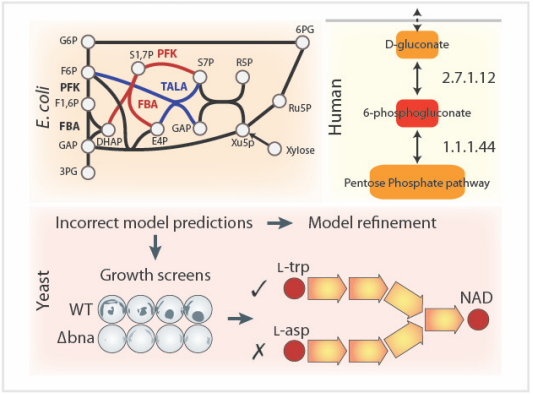
B. Comparing inputs and outputs



C. Iterative improvement



D. Gap filling approaches



E. Discovery of regulatory interactions

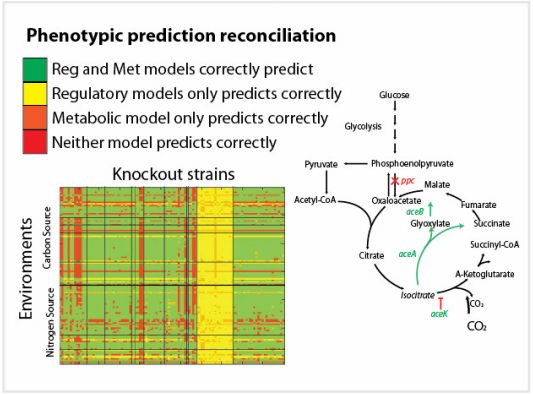


Figure 1.4. Quantitative phenotype prediction using optimization
A) Quantitative phenotype prediction can be conceptualized as an iterative workflow. First, hypothesized biological constraints and objective are formulated mathematically, and computational optimization is used to determine optimal phenotypic states (see Section 2). The predicted phenotypic states can then be compared to experimental measurements to identify where predictions are consistent. When consistent, the hypothesized evolutionary objective and constraints are validated. When inconsistent, laboratory evolution can be used to gain further insight as to why the computed and measured states differ. Examples of validation of quantitative phenotypes are detailed in **4B** and further hypotheses derived from incorrect predictions are detailed in **4C**. **B)** The generic workflow in 4A has been successfully applied to several classes of phenotypes. i) The ratios of nutrient utilization can be predicted by maximizing biomass flux across different substrate uptake bounds (Edwards et al., 2001); the resulting surface is referred to as a phenotypic phase plane, and is a standard function in the Cobra toolbox. ii) Central carbon metabolism fluxes can be predicted. For some organisms, much of the variability in flux can be attributed to biomass flux maximization (Schuetz et al., 2012). Specific deviations have led to new understanding of relevant biological constraints and objectives (see **Figure 1.4C**). iii) The ratio of organism abundances and nutrient exchanges can be predicted for both natural and synthetic communities when appropriate constraints limiting nutrient exchange are applied. Note that one important feature of quantitative phenotype predictions is that optimal flux solutions are often not unique. That is, there are multiple equivalent phenotypic states that can achieve the same objective value. To address this, flux variability analysis (FVA) (Mahadevan and Schilling, 2003b) can be used to identify the ranges of possible fluxes. It should be noted that non-uniqueness is not necessarily a handicap of COBRA as biological evolution can come up with alternate solutions (Fong et al., 2005). C) Inconsistencies with model predictions have led to the appreciation of new constraints and objectives underlying cellular phenotypes. i) Inconsistent predictions in by-product secretion have led to the hypothesis that membrane space limits membrane protein abundance and metabolic flux (Zhuang et al., 2011b). This constraint can be added by bounding fluxes across the membrane. ii) The range of metabolic fluxes observed across different environments have led to the understanding that fluxes can be understood as simultaneously satisfying multiple competing objectives, such as growth and cellular maintenance. Multi-objective optimization algorithms find solutions that maximize multiple competing objectives. **D)** Accurate prediction of quantitative phenotypes has led to prospective design of biological functions. A number of algorithms have been developed that predict genetic and/or environmental perturbations required to achieve a bioengineering objective. Relevant bioengineering objectives have included biosensing, bioremediation, bioproduction, the creation of synthetic ecologies, and the intracellular production of reaction oxygen species (ROS) to potentiate antibiotic effects (Brynildsen et al., 2013b). For example, growth-coupled production of metabolites can be computed with the OptKnock algorithm (Burgard et al., 2003).

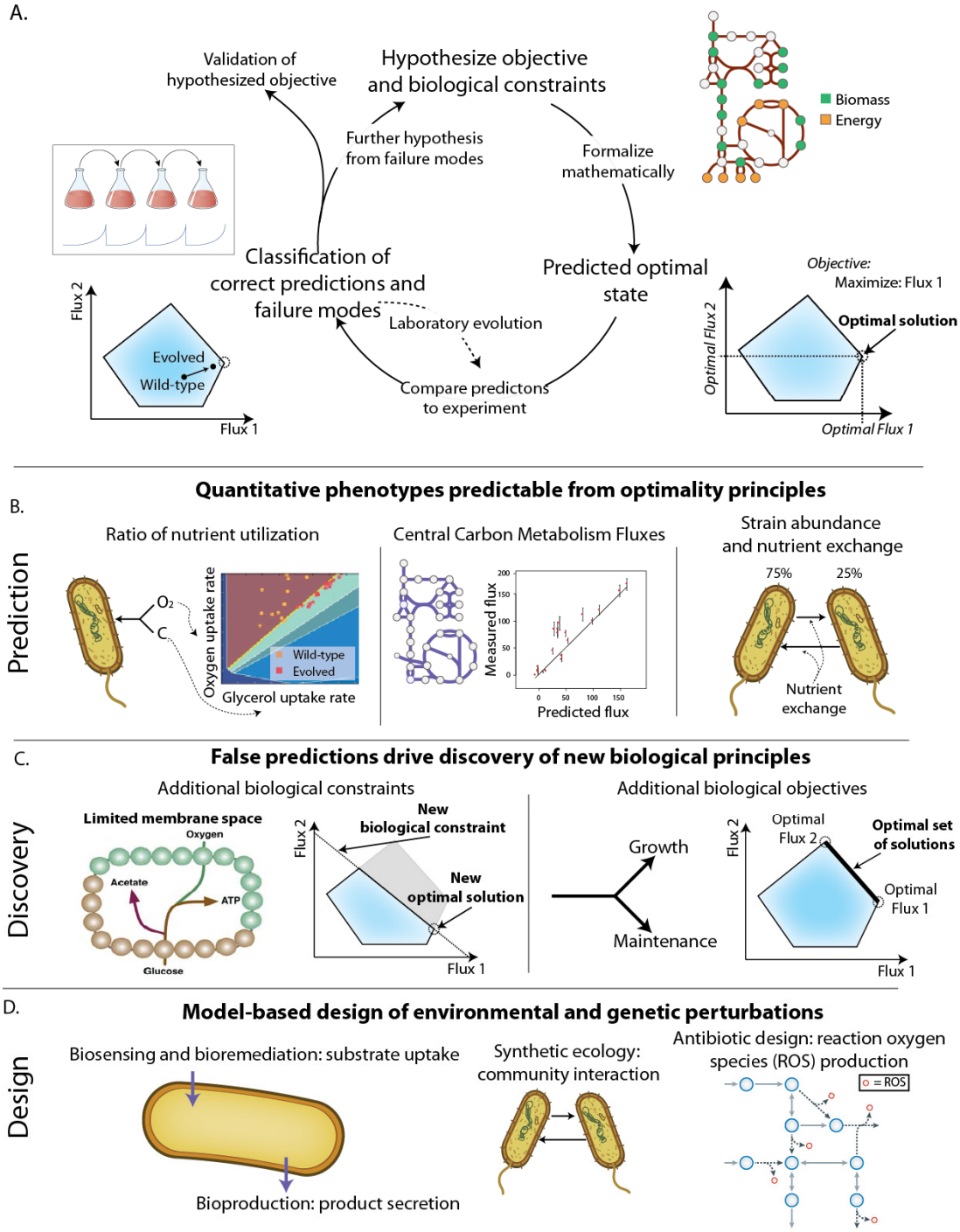

A.
Validation of hypothesized objective
Hypothesize objective and biological constraints
Further hypothesis from failure modes
Formalize mathematically
Biomass
Energy
Classification of correct predictions and failure modes
Predicted optimal state
Laboratory evolution
Compare predictons to experiment
Flux 2
Flux 1
Evolved
Wild-type
Objective:
Maximize: Flux 1
Optimal solution
Optimal Flux 2
Optimal Flux 1
B.
Quantitative phenotypes predictable from optimality principles
Prediction
Ratio of nutrient utilization
Central Carbon Metabolism Fluxes
Strain abundance and nutrient exchange
O2
C
Oxygen uptake rate
Glycerol uptake rate
Wild-type
Evolved
Measured flux
Predicted flux
75%
25%
Nutrient exchange
C.
False predictions drive discovery of new biological principles
Discovery
Additional biological constraints
Additional biological objectives
Limited membrane space
Oxygen
Acetate
ATP
Glucose
New biological constraint
New optimal solution
Flux 2
Flux 1
Growth
Maintenance
Optimal Flux 2
Optimal set of solutions
Optimal Flux 1
D.
Model-based design of environmental and genetic perturbations
Design
Biosensing and bioremediation: substrate uptake
Bioproduction: product secretion
Synthetic ecology: community interaction
Antibiotic design: reaction oxygen species (ROS) production
= ROS

Figure 1.5. Data integration and exploration of possible cellular phenotypes
A) The general workflow for multi-omic data integration begins with the conversion of the experimental data into model constraints (see **Figure 1.5B**). This procedure results in cell-type (e.g. neuron, macrophage) and condition-specific (e.g. healthy vs. diseased) models that represent the metabolic capabilities of those specific cells (see **Figure 1.5C**). Several computational procedures can then be used to explore the metabolic capabilities and determine achievable phenotypes systematically (see **Figure 1.5D**). Evaluation of these phenotypic capabilities and comparison of different cells or environments leads to identification of their molecular differences (see **Figure 1.5E**). Additionally, if the original experimental data cannot precisely distinguish between certain metabolic states, additional targeted experiments can be designed and integrated as further constraints. **B)** Numerous data types can be integrated into metabolic models. Some directly affects model structure and variables (e.g. growth rate, biomass composition, exchange fluxes, internal fluxes and reaction directionality). Standard processing of these data types allows for integration into the model. Other data types affect metabolic fluxes more indirectly. As such, different computational methods exist for formulating the appropriate constraints (Error! Reference source not found.). **C)** Experimental data is integrated to construct cell-type and/or condition-specific models. These models represent the metabolic capabilities in a certain state, and are then used for further inquiry (see Figures 5D,E). Specific algorithms for building cell-type specific models from gene expression data include MBA (Jerby et al., 2010) and GIMME (Becker and Palsson, 2008). **D)** After adding constraints to the model, computational procedures are used to assess the implication of the experimental data on metabolic fluxes. The two main methods for querying the consequences of the measured data on a cell's phenotypes are flux variability analysis (FVA) and Markov-chain Monte-Carlo (MCMC) sampling. These are both standard functions in the Cobra toolbox. i) FVA determines the maximum and minimum values of all metabolic fluxes. ii) MCMC sampling randomly samples feasible metabolic flux vectors (usually resulting in tens to hundreds of thousands of flux vectors). These sampled flux vectors can then be used to derive the distribution of possible flux values for a given metabolic reaction. One common pitfall to be aware of in FVA and sampling is the presence of flux cycles (or loops) in the metabolic network. Cycles occur when a set of reactions perfectly balance each other, resulting in no net metabolite production or consumption. In actuality, these cycles are not thermodynamically feasible (Noor et al., 2012) and should be ignored or removed with computational approaches (Schellenberger et al., 2011a). **E)** Often a comparative approach is employed in which experimental data from two conditions are used to generate two condition-specific models. Then, the achievable phenotypes of the two states are compared (e.g. though MCMC sampling, see **Figure 1.5D**).

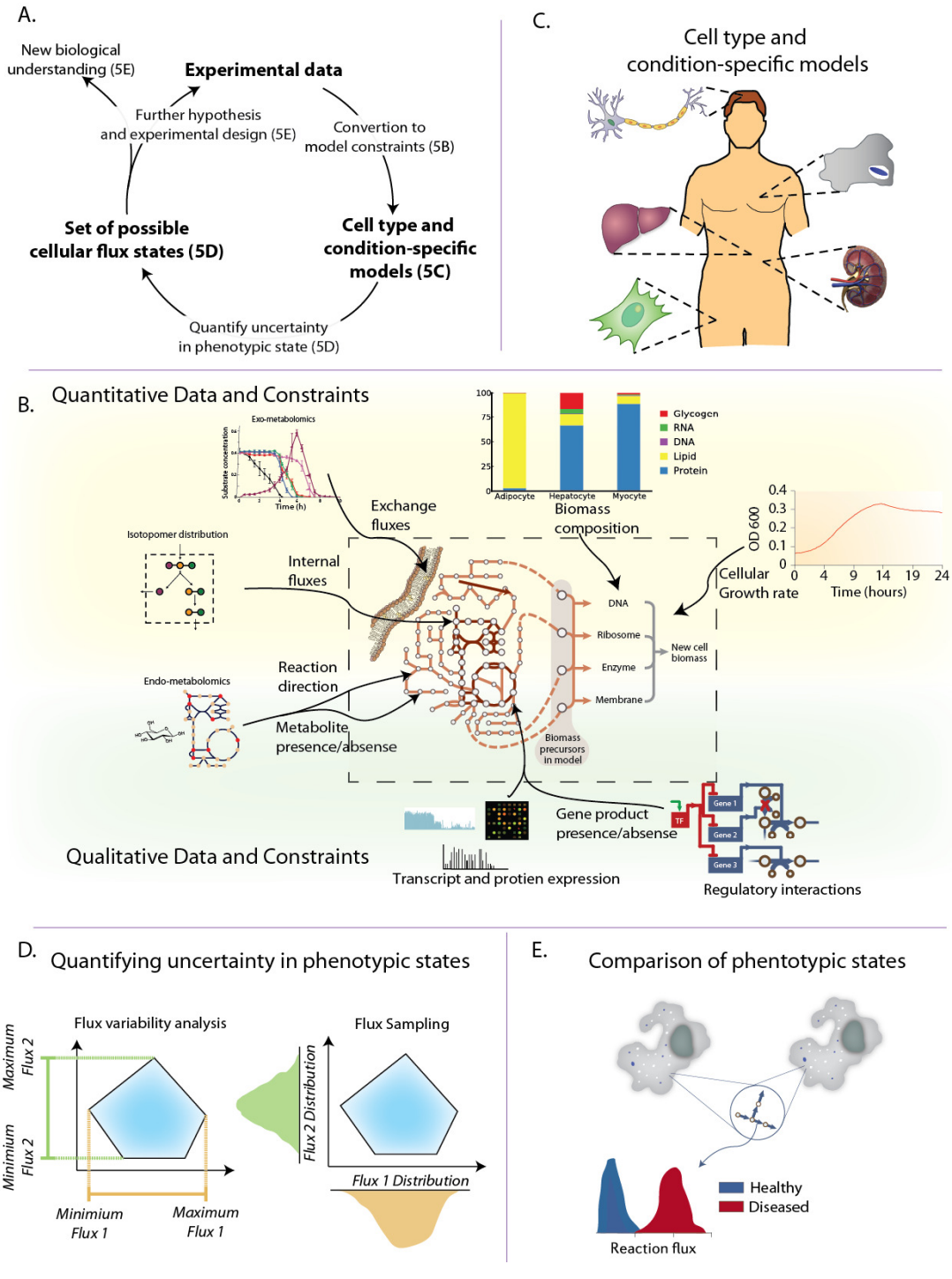
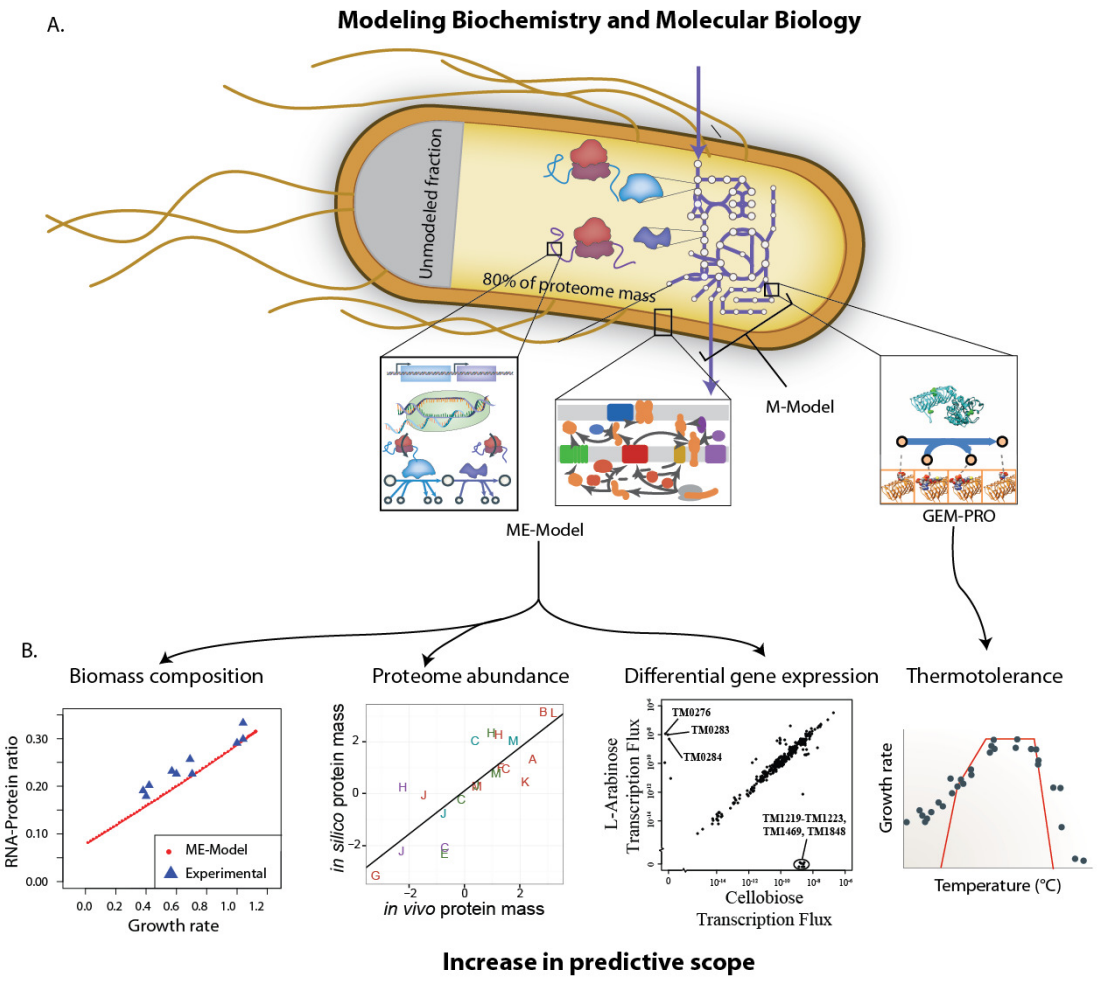

A.
New biological understanding (5E)
Experimental data
Further hypothesis and experimental design (5E)
Convertion to model constraints (5B)
Set of possible cellular flux states (5D)
Cell type and condition-specific models (5C)
Quantify uncertainty in phenotypic state (5D)
C.
Cell type and condition-specific models
B.
Quantitative Data and Constraints
Exo-metabolomics
Substrate concentration
Time (h)
Exchange fluxes
Isotopomer distribution
Internal fluxes
Endo-metabolomics
Reaction direction
Metabolite presence/absense
Glycogen
RNA
DNA
Lipid
Protein
Adipocyte
Hepatocyte
Myocyte
Biomass composition
OD 600
Cellular Growth rate
Time (hours)
DNA
Ribosome
Enzyme
Membrane
New cell biomass
Biomass precursors in model
Gene product presence/absense
TF
Gene 1
Gene 2
Gene 3
Qualitative Data and Constraints
Transcript and protien expression
Regulatory interactions
D.
Quantifying uncertainty in phenotypic states
Flux variability analysis
Maximum Flux 2
Minimium Flux 2
Minimium Flux 1
Maximum Flux 1
Flux Sampling
Flux 2 Distribution
Flux 1 Distribution
E.
Comparison of phentotypic states
Healthy
Diseased
Reaction flux

Figure 1.6. Expansion of genome-scale models to encompass molecular biology
A) Metabolic models have been expanded to encompass the processes of proteome synthesis and localization as well as data on protein structures. Models including protein synthesis and localization are referred to as ME-Models, which stands for metabolism and gene expression. GEM-PRO refers to genome-scale models integrated with protein structures. For GEM-PRO, a combination of structural data directly references the GPRs in the metabolic reconstruction; structures can be obtained from experimental databases or homology modeling. The *E. coli* ME-Model mechanistically accounts for ~80% of the proteome mass in conditions of exponential growth and 100% of other major cell constituents (DNA, RNA, cell wall, lipids, etc). **B)** Addition of cellular processes vastly increases the predictive scope of models. ME-Models can predict biomass composition, abundances of protein across subsystems, and differential gene expression in certain environmental shifts (in addition to the predictions possible with M-Models); like FBA these were predicted by assuming growth maximization as an evolutionary objective, though the specific optimization algorithm differs due to the addition of coupling constraints. GEM-PRO has been used to predict the metabolic bottlenecks and growth defects of changes in temperature on protein stability and catalysis; protein stability is predicted with structural bioinformatics methods and then used to limit the catalyzed metabolic flux. The uses of these integrated models are just beginning to be explored.


A.
Modeling Biochemistry and Molecular Biology
Unmodeled fraction
80% of proteome mass
M-Model
ME-Model
GEM-PRO
B.
Biomass composition
Proteome abundance
Differential gene expression
Thermotolerance
RNA-Protein ratio
Growth rate
ME-Model
Experimental
in silico protein mass
in vivo protein mass
L-Arabinose Transcription Flux
Cellobiose Transcription Flux
TM0276
TM0283
TM0284
TM1219-TM1223, TM1469, TM1848
Growth rate
Temperature (°C)
Increase in predictive scope

Chapter 2 Comparative Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments

2.1 Abstract

Genome-scale models (GEMs) of metabolism were constructed for 55 fully sequenced *Escherichia coli* and *Shigella* strains. The GEMs enable a systems approach to characterizing the pan and core metabolic capabilities of the *Escherichia coli* species. The majority of pan metabolic content was found to consist of alternate catabolic pathways for unique nutrient sources. The GEMs were then used to systematically analyze growth capabilities in over 650 different growth-supporting environments. The results show that unique strain-specific metabolic capabilities correspond to pathotypes and environmental niches. Twelve of the GEMs were used to predict growth on six differentiating nutrients and the predictions were found to agree with 80% of experimental outcomes. Additionally, GEMs were used to predict strain-specific auxotrophies. Twelve of the strains modeled were predicted to be auxotrophic for vitamins niacin (vitamin B3), thiamin (vitamin B1) or folate (vitamin B9). Six of the strains modeled have lost biosynthetic pathways for essential amino acids methionine, tryptophan or leucine. Genome-scale analysis of multiple strains of a species can thus be used to define the metabolic essence of a microbial species and delineate growth differences that shed light on the adaptation process to a particular micro-environment.

2.2 Introduction

Over the past decade, the *E. coli* K-12 MG1655 strain has been used extensively as a model organism for research on microbial metabolic systems biology. However, the increasing availability of genomic sequences for other *E. coli* strains suggests that this non-pathogenic laboratory strain's genes are a small part of the genomic diversity in the *E. coli* species. For instance, the *E. coli* O157:H7 EDL933 strain responsible for worldwide outbreaks of hemorrhagic colitis has one million more base pairs of DNA than

K-12 MG1655 (approximately 20% larger) (Hayashi et al., 2001, Perna et al., 2001, Canchaya et al., 2003). Multiple genomic sequences have defined a set of genes that is common to all *E. coli* strains (i.e., a 'core' genome), and it has been determined that they represent a small fraction of the entire *E. coli* gene pool. The growing availability of whole genome sequences for *E. coli* strains thus brings into focus the question 'what is a strain and what is a species?'

A recent study of 20 *E. coli* strains found that a large fraction of the shared genomic elements with known function are related to metabolism (Touchon et al., 2009). Therefore, it is important to characterize the genes that encode a core set of metabolic capabilities to understand their effect on cellular functions, as they constitute a common denominator that can be used to define the core metabolic potential of the *E. coli* species. Metabolic network reconstructions have proven to be powerful tools to probe the genomic diversity of metabolism between organisms (Liao et al., 2011b, Baumler et al., 2011, Archer et al., 2011, Yoon et al., 2012, Charusanti et al., 2011, Thiele et al., 2011). As useful as genome annotation is, it does not provide an understanding of the integrated function of gene products to produce phenotypic states. K-12 MG1655 was the first *E. coli* strain to have its genome entirely sequenced (Blattner et al., 1997). A first genome-scale metabolic reconstruction was completed for this strain three years later (Edwards and Palsson, 2000a). Since then, the reconstruction of MG1655 has undergone a series of expansions in the intervening 13 years as more information about the genome and its annotation has become available (Reed et al., 2003, Feist et al., 2007, Orth et al., 2011). The most recent reconstruction, iJO1366 (Orth et al., 2011), accounts for 1366 genes (39% of functionally annotated genes on the genome) and their gene products. The genome-scale metabolic reconstruction for *E. coli* K-12 MG1655 is

the most complete metabolic reconstruction available to date (McCloskey et al., 2013c, Feist and Palsson, 2008b). However, as more *E. coli* genomic sequences have become available, it has become clear that *E. coli* K-12 MG1655 only partially represents the *E. coli* species. Thus, it is important to construct GEMs for other *E. coli* strains because of this species' importance to human health, basic microbiological science and industrial biotechnology (Lee, 2009).

The goal of this study was to construct GEMs for all *E. coli* strains with fully sequenced genomes and thus, for the first time, to reconstruct the metabolic network for an entire species and its strain-specific variants. *Shigella* strains were included on the basis of 16S ribosomal profiling experiments that classify *Shigella* strains as members of the *E. coli* species (Pupo et al., 2000), despite the historical distinction of having their own genus. Therefore, the formulated GEMs span commensal strains as well as both intestinal and extra-intestinal pathogenic strains of *E. coli* allowing for a comprehensive analysis of the representative metabolic capabilities of the *E. coli* species.

2.3 Results

2.3.1 Characteristics of E. coli core and pan metabolic content

A set of 55 *E. coli* genome-scale reconstructions was built and used to compared gene, reaction, and metabolite content between strains (**Dataset S1**). The content shared among all reconstructions thereby defines the 'core' metabolic capabilities among all the strains. Similarly, the metabolic capabilities of all the strains were combined to define the full set that encompasses all models and thereby define the 'pan' metabolic capabilities among all the strains. By analogy to mathematical set theory, the core metabolic content is the intersection of the gene, reaction, and metabolite content

of all 55 models while the pan metabolic content is the union of these features among the models (

Figure 2.1A).

The size and content of the core metabolic content characterizes the metabolic foundation of *E. coli* as a species. The core model has 965 metabolic genes that catalyze 1,773 reactions using 1,665 metabolites. The most highly conserved subsystems were lipid metabolism, cell wall, membrane and envelope metabolism, nucleotide metabolism and cofactor and prosthetic group metabolism. Reactions involved in lipid metabolism, cell wall/membrane/envelope metabolism and cofactor and prosthetic group biosynthesis were highly represented (>80%) in the core reactome. Most of these reactions synthesize essential components such as vitamins and cofactors like riboflavin, coenzyme A and biotin, as well as quinones and isoprenoids. In contrast, only 36% of carbohydrate metabolism reactions were part of the core reactome. These reactions were comprised of central metabolism reactions including anaplerotic reactions, the citric acid cycle, glycolysis/gluconeogenesis and the pentose phosphate pathway.

The pan metabolic capabilities are comprised of the total number of different reactions found in all strains and are thus an indicator of the full metabolic capabilities within a species. The *E. coli* pan reconstruction content contains 1460 metabolic genes, 2501 reactions and 2043 metabolites. About 64% of reactions in carbohydrate metabolism were part of the pan reactome, the largest group (

Figure 2.1B). A majority of these reactions are involved in alternate carbon source metabolism. Cell wall and membrane envelope metabolism accounted for 18% of reactions in the pan reactome. These reactions account for a major phenotypic

distinction between *E. coli* strain's serogroup, in particular the O-antigen (DebRoy et al., 2011). Also, 30% of amino acid metabolism reactions are part of the pan reactome.

2.3.2 The ability to catabolize different nutrient sources distinguishes metabolic models of E. coli strains

The conversion of static metabolic network reconstructions into computable mathematical models allows computation of phenotypes based on the content of each reconstruction. Thus, the 55 strain-specific reconstructed networks were converted into genome-scale metabolic models (GEMs) that allow for the simulation of phenotype (Feist et al., 2009). This set of GEMs allows for a meaningful interpretation of the content of each reconstruction and allows for the prediction of a strain's micro-environmental and ecological niche.

Reactions belonging to the alternate carbon metabolism subsystem made up the majority of reactions in the pan reactome (

Figure 2.1B). Thus, it was hypothesized that these capabilities may reflect functional differences in the ability of different strains of *E. coli* to adapt to different nutritional environments. To test this hypothesis, growth was simulated *in silico* for all 55 *E. coli* and *Shigella* GEMs on minimal media in 654 growth conditions. The conditions were composed of all sole growth supporting carbon, nitrogen, phosphorous and sulfur sources in both aerobic and anaerobic environments (**Figure 2.2**).

In contrast to the *E. coli* GEMs, the *Shigella* GEMs displayed a large loss of catabolic capabilities across the 654 growth conditions. This computational result supports evidence showing that *Shigella* strains have lost catabolic pathways for many nutrient sources (Bliven and Maurelli, 2012). Models of *Shigella* strains completely lost the capability to sustain growth on nutrient sources for which more than 90% of *E. coli*

models had growth capabilities. Some of these nutrients include D-alantoin, D-malate, and xanthine as carbon sources as well as inosine as a nitrogen source. Furthermore, only 1 out of the 8 *Shigella* strain models (13%) was able to sustain growth on choline or L-fucose, two carbon sources that most *E. coli* strain models examined were predicted to catabolize.

2.3.3 A set of substrates differentiate pathogenic strains from commensal (non-pathogenic) strains

Based on simulated growth phenotypes, we observed a general separation of commensal strains from both Extra-intestinal Pathogenic *E. coli* (ExPec) and Intestinal Pathogenic *E. coli* (InPec) strains of *E. coli* suggesting that a classification schema of strains based on metabolic capabilities is possible (**Figure 2.3**). Common lab strains of *E. coli* such as *E. coli* K-12 MG1655 are non-pathogenic, commensal strains. As a first step towards establishing such a schema, the separation between ExPec and commensal strain models was examined. A Fisher's exact test was used to establish that models of ExPec strains exhibited a statistically significant capability to catabolize four unique compounds with a p-value of less than 0.05 (**Table 2.1A**).

Most models of strains widely regarded as safe lab strains such as K-12 strains, BW2952, and DH1 were unable to grow on a unique subset of nutrients. Notably, N-Acetyl-D-galactosamine supported growth in 100% of ExPec strain models compared to 67% of commensal strain models ($p=3.9x10^{-2}$). Additionally, several commensal strain models exhibited a statistically significant overrepresentation of catabolic pathways for 13 nutrient sources (**Table 2.1B**). For example, fructoselysine and psiscoselysine share the same catabolic pathway and were not catabolizable by any of the ExPec models; however, 89% ($p=2.2x10^{-6}$) of the intestinal strain models could utilize fructoselysine or

psiscoselysine as a sole carbon source. Fructoselysine is poorly digested in the human small intestine and little is excreted, hinting that the majority of dietary fructoselysine may be digested by the intestinal microbiota (Erbersdobler and Faist, 2001). Further, 4-hydroxyphenylacetate, an aromatic compound, was catabolized as a sole carbon source for 55% of commensal strain models compared to only 9% of ExPec strain models ($p=1.3x10^{-2}$). Hydroxyphenylacetic acids are produced by bacterial fermentation of short chain peptides and amino acids in the human large intestine (Smith and Macfarlane, 1996). 4-hydroxyphenylacetate undergoes eight different enzymatic reactions before being converted to pyruvate and succinate-semialdehyde that can then be converted to succinate and enter the TCA cycle (Prieto et al., 1996).

Next, the models of strains known to reside intestinally were compared to investigate differences between commensal and InPec strains. Models of InPec strains displayed an advantage in their ability to support growth on seven unique carbon and nitrogen sources (**Table 2.2A**). Some of the substrates had unique enrichment specifically among the enterohemmorhagic (EHEC) strains, responsible for worldwide cases of diarrhea and hemolytic uremic syndrome (HUS) (Karch et al., 2005). Sucrose supported growth for 65% of the InPec strains including 100% of EHEC strains compared to only 33% of commensal strains ($p=5.0x10^{-2}$). Also, consistent with other reports (Nakano et al., 2001), urease activity was present in EHEC strain models only. Urea supported growth as a sole nitrogen source for 47% of InPec strain models, including 100% of the EHEC models compared with 0% of commensal strain models ($p=1.0x10^{-3}$). Urease degrades urea into CO_2 and NH_4 and therefore may provide an additional source of nitrogen for cells in nitrogen-limited environments.

In contrast to InPec strains, commensal strains displayed an advantage in their ability to degrade eleven unique carbon and nitrogen sources (**Table 2B**). The short chain fatty acids (SCFAs) acetoacetate and butyrate were found to support growth as sole carbon sources for 78% of commensal strain models compared to 47% of InPec models ($p=5.0x10^{-2}$). Notably, none of the EHEC strain models were able to catabolize either of these two compounds and only 13% of *Shigella* models were able to utilize them as a sole source of carbon.

2.3.4 Metabolic models combined with gap-filling methods facilitate investigation into the genetic basis of strain-specific auxotrophies

In addition to investigating growth-supporting nutrients, GEMs can also be used to examine the genetic bases of strain-specific auxotrophies. Twelve of the 55 reconstructed GEMs were unable to generate essential biomass components from glucose M9 minimal media without addition of growth-supporting compounds to the *in silico* media. The SMILEY algorithm, a method to fill gaps in metabolic networks (Reed et al., 2006a), was used to examine the genetic bases of these model auxotrophies (**Figure 2.4**). Based on this analysis, six of the eight *Shigella* strains exhibited an auxotrophy for niacin (vitamin B3) *in silico*. These simulation results are consistent with literature data indicating that many strains of *Shigella*, including *Shigella sonnei* Ss046 and *Shigella boydii* sb227, are unable to grow without addition of niacin to M9 minimal media with glucose (Prunier et al., 2007a). Gap analysis attributes this auxotrophy to the lack of L-aspartate oxidase activity, encoded by the gene *nadB*, in the nicotinic acid biosynthesis pathway. A bioinformatic analysis of *nadB* suggests that it is a pseudogene due to numerous non-synonymous mutations compared to the sequence of *nadB* in *E. coli* K-12.

Two additional examples of identifying and confirming strain-specific auxotrophies involved models for *Shigella flexneri* 2a str 301, (auxotrophic for methionine) and *E. coli* strain DH10B (auxotrophic for leucine). Gap analysis of *S. flexneri* 2a str 301 suggested that the auxotrophy is due to the absence of homoserine O-trans-succinylase, encoded by *metA* in *E. coli* K-12 MG1655 (b4013). A bioinformatics analysis suggested that *metA* is a pseudogene in *S. flexneri* 2a str 301 due to single base pair deletion, causing a frameshift mutation at amino acid position 212/310 and hence premature termination of the full length protein. Both of these observations were confirmed with literature evidence. Specifically, *S. flexneri* strains are known to require methionine (Zagaglia et al., 1991) in minimal media and *E. coli* DH10B is known to require leucine due to a deletion of the *leuABCD* operon (Durfee et al., 2008).

2.3.5 Experimental validation of unique nutrients shows high model accuracy

To assess the accuracy of *in silico* growth simulations, 12 of the 55 reconstructed strains were screened for growth on six carbon sources. Growth was estimated by optical density 48 hours after inoculation. OD values of >0.08 were considered growth. The 12 strains consisted of 3 ExPec strains, 3 InPec strains, 5 commensal strains and 1 *Shigella* strain; thereby spanning the pathotypes discussed above. Six carbon sources were selected based on their predicted ability to classify strains according to different pathotypes. In other words, each of these six substrates was expected to support growth of certain strains but not others. The models are validated by true positive and true negative results that highlight cases where the models are in agreement with experimental results. In contrast, false positive and false negative cases indicate potential errors or gaps in the models (Feist and Palsson, 2008b) (**Figure 2.5A**). The

experimental results showed a high level of accuracy with 80% of GEM predictions agreeing with experiments.

Comparison of *in silico* and experimental results revealed complete agreement for two different carbon sources: acetoacetate and deoxyribose were predicted correctly with 100% accuracy (8 and 2 true positives, as well as 4 and 10 true negatives, respectively) across all 12 strains (**Figure 2.5B**). Acetoacetate is transported into the cell via a proton symporter encoded by *atoE*. Growth of *E. coli* on short-chain fatty acids, such as acetoacetate, requires activation of the acid to its respective thioester (Jenkins and Nunn, 1987). For acetoacetate, this activation is catalyzed by acetoacetyl-CoA transferase encoded by *atoA* and *atoD* that form a four unit enzymatic complex. The four strains lacking these two genes were correctly predicted not to grow on acetoacetate as a sole carbon source. Deoxyribose is the second compound predicted with 100% accuracy. It is transported into the cell via a proton symporter encoded by *deoP*. Deoxyribose is then phosphorylated to deoxyribose-5-phosphate by deoxyribose kinase, encoded by *deoK (Bernier-Febreau et al., 2004)*. Finally, deoxyribose-5-phosphate is converted into acetaldehyde and glyceraldehyde-3-phosphate by deoxyribose-phosphate aldolase encoded by *deoC*. Only two of the strains tested, *E. coli* O42 and *E. coli* CFTO73, possessed these genes and were correctly predicted by the models to be able to grow on deoxyribose as a sole carbon source. These cases validate the approach and demonstrate high accuracy to discriminate between different strains' capabilities.

A single pathway for two aromatic phenyl compounds, phenylacetaldehyde and phenylethylamine, was responsible for significant differences between model predictions and experimental results. The growth predictions for phenylacetaldehyde and

phenylethylamine (tested individually) exhibited identical profiles of 7 false negatives and only 5 true positives during *in vivo* growth screens (42% accuracy). These two compounds share a catabolic pathway whereby phenylethylamine is converted to phenylacetaldehyde by phenylethylamine oxidase encoded by *tynA* (Diaz et al., 2001). Phenylacetaldehyde dehydrogenase, encoded for by *feaB,* then converts phenylacetaldehyde to phenylacetic acid. The acid is subsequently converted to phenylacetyl-CoA by phenylacetate-CoA ligase encoded by *paaK*. The genes coding for enzymes that catalyze these reactions in *E. coli* K-12 MG1655 had very low identity (<40%) at the amino acid level to genes in strains that proved capable of utilizing these substrates as sole sources of carbon. Therefore, the growth experiments indicate that either these low identity enzymes are carrying out the activity or there is an alternate pathway for catabolism of these two compounds.

2.4 Discussion

Genome-scale models (GEMs) of metabolism are powerful tools that can be deployed to investigate similarities and differences between strains of the same species. Unique GEMs for 55 different *E. coli* strains were constructed and used to: 1) compare and contrast core and pan metabolic capabilities within the *E. coli* species; 2) determine functional differences between strains by computing growth phenotypes on over 650 different nutrients both aerobically and anaerobically; and 3) explore the genetic basis behind strain-specific auxotrophies. These computational classifications and studies were fortified by performing *in vivo* screens of select discriminating compounds and strains resulting in a high level of accuracy (80%).

The majority of reactions found in pan metabolism fell into the metabolic subsystem of alternate carbon metabolism. It was hypothesized that these differences

give different strains advantages in preferred micro-environmental niches. A clustering analysis based on computed metabolic phenotypes clearly distinguished *E. coli* strains from *Shigella* and largely separated *E. coli* strains known to exhibit a commensal intestinal lifestyle from those known to exhibit both intestinal and extra-intestinal pathogenic lifestyles. This separation was based solely on the catabolic capabilities of different strains for unique nutrient sources. A major distinction that appeared was the capability to degrade fructoselysine and psiscoselysine indicating that these compounds may be a defining feature of intestinal *E. coli* strains. This pathway is missing from all extra-intestinal *E. coli* strains investigated. One possible mechanism that explains this feature begins with the observation that fructoselysine is poorly digested in the small intestine and absorption occurs only through diffusion (Erbersdobler and Faist, 2001). As a result, fructoselysine moves to the large intestine where it is present in abundance. However, excretory levels of fructoselysine are low, thus it has been postulated that the intestinal microbiota, including *E. coli*, ferment almost all dietary fructoselysine. *E. coli* K-12 MG1655 has been shown to sustain growth on fructoselysine as a sole carbon source in anaerobic environments (Erbersdobler and Faist, 2001) and 88% of the intestinal strains modeled are predicted to be capable of utilizing fructoselysine or psiscoselysine as a sole carbon source.

Another example of strain discrimination was the aromatic compound 4-hydroxyphenylacetic acid that was catabolized by 55% of commensal intestinal strain models compared to 9% of extra-intestinal strain models. Hydroxyphenylacetic acids are one of several classes of aromatic compounds produced by bacterial fermentation of short chain peptides and amino acids in the human large intestine (Smith and Macfarlane, 1996). Specifically, 3- and 4-hydroxyphenylacetic acid have been identified

as products of tyrosine fermentation (Diaz et al., 2001) by the diverse colonic microbiota. Therefore, these compounds are likely present at high levels in the intestine. Thus utilization of 4-hydroxyphenylacetic acid as a sole carbon source may provide a competitive advantage over other strains of *E. coli* in the gut.

In addition to unique growth capabilities, the GEMs are also able to reliably predict strain-specific auxotrophies. This ability is important as auxotrophies often indicate cases of antagonistic pleiotrophy whereby ancestral traits that interfere with virulence are lost to a newly evolved pathogen. Traits absent in pathogenic strains of a species but commonly expressed in commensal ancestors are strong candidates for pathoadaptive mutations. Evidence of this model of pathogen evolution was first provided by *Shigella* and *E. coli*. Widespread niacin auxotrophies in *Shigella* strains were identified due to disruption of *nadA* and *nadB* genes that code for the enzyme complex that converts L-aspartate to quinolate, a precursor to NAD synthesis. This finding is validated by previous literature confirming that quinolate inhibits invasion and cell-to-cell spread of *Shigella flexneri* 5a. Reintroduction of functional copies of *nadA* and *nadB* into this strain restored the ability to synthesize quinolate but also resulted in strong attenuation of virulence in this strain (Prunier et al., 2007b). Therefore, several of the additional auxotrophies identified for other vitamins folate and thiamin as well as amino acids leucine, methionine and tryptophan may indicate further cases of antagonistic pleiotrophy. Future studies could explore the impact of these auxotrophies on virulence in each strain to potentially elucidate new pathoadaptive mutations.

Growth experiments were performed for six carbon sources tested on twelve different strains to evaluate the predictability of the developed models. The overall accuracy of the models was 80%, a level that is in line with predecessor models (Orth et

al., 2011, Feist et al., 2007, Reed et al., 2003, Edwards and Palsson, 2000a). This high level of accuracy is notable because the substrates tested were selected due to their ability to differentiate amongst strains, making them some of the most difficult compounds to correctly predict. Three of the strains, 8739, HS, and MG1655 had 100% predictive accuracy. These are all safe, commensal lab strains which likely contributed to them having better genome annotations and subsequently more accurate model predictions. Cases where the models are incorrect provide opportunities for biological discovery. False positives represent missing context-specific information in a GEM. These occur when model predicted growth on a compound disagrees with the lack of growth observed experimentally. For example, growth on D-malate was a false positive for two models of *E. coli* strains. Even though both strains have a gene that has high identity to the D-malate decarboxylating oxidoreductase enzyme, encoded by *yeaU*, they were unable to grow on this compound. This could indicate a case where expression of this enzyme is transcriptionally repressed. Adaptive laboratory evolution of these two strains on D-malate may relieve the transcriptional repression of *yeaU* and lead to identification of novel regulators involved in controlling the catabolism of this compound (Lee and Palsson, 2010).

In contrast to false positives, false negatives occur when comparing *in silico* and *in vitro* data to identify missing content in a GEM. Growth predictions for phenylacetaldehyde and phenylethylamine consisted of seven false negatives. These two compounds share the same catabolic pathway. The three genes catalyzing reactions in this pathway for *E. coli* K-12 MG1655 had very low identity (<40%) to genes in strains that proved capable of utilizing these substrates as sole sources of carbon. Therefore, the growth experiments indicate that either a particular domain on these low

identity enzymes is carrying out the activity or there is an alternate pathway for catabolism of these two compounds. Further characterization and studies on gene knock-outs for this pathway in each strain could lead to identification of a new alternate pathway for phenylacetaldehyde and phenylethylamine catabolism.

The work presented here shows that strain-specific models of *E. coli* can guide further studies regarding the advantages conferred by unique nutrients to *E. coli* strains in different niches. Additionally, the models reliably predict strain-specific auxotrophies documented in the literature as well as novel auxotrophies that offer a strong case for future study. Taken together, this study represents a step towards the definition of a bacterial species based on common metabolic capabilities and its strains based on niche-specific growth capabilities. In addition to this fundamental advance, the niche-specific characteristics provide a basis for understanding strain and species-specific pathogenesis. Similar studies of diverse strains for species beyond *E. coli* will further define the concept of a species. Ultimately, this understanding can be leveraged to formulate strain- and species-specific drug development and therapeutic approaches.

2.5 Materials and Methods

2.5.1 Strain specific model reconstruction

All genomes were re-annotated using the RAST server (Aziz et al., 2012, Aziz et al., 2008). Re-annotation led to 600 new genes being annotated. Genes that were annotated as pseudogenes in the original NCBI annotation were treated as pseudogenes and the enzymatic function of the proteins they were removed from the final models. A total of 567 metabolic pseudogenes were identified. Gene sequences from six metabolic models for *E. coli* K-12 MG1655 (Orth et al., 2011), *Salmonella typhimurium* LT2 (Thiele et al., 2011), *Klebsiella pneumoniae* MGH 78578 (Liao et al.,

2011b) and *Yersinia pestis* CO92 (Charusanti et al., 2011), *E. coli* W (Archer et al., 2011) and *E. coli* B REL606 (Yoon et al., 2012) were used for identifying orthologs. The SEED Corresponding Genes tool was used to identify orthologs in each strain of *E. coli* (Aziz et al., 2012). This tool identifies best bi-directional hits (BBH) and accounts for gene context (Binter et al., 2012). A 70% percentage identity (PID) cutoff was used for assigning orthologs. Genes that were missing orthologs in the original models were deleted from the model for the target strain. Additional reaction content was added from ModelSEED (Henry et al., 2010c) , KEGG (Kanehisa et al., 2012, Kanehisa and Goto, 2000) and BIOCYC (Caspi et al., 2008). All reactions added were manually curated according to published protocol (Thiele and Palsson, 2010b). MetaNetX (Ganter et al., 2013) was used to standardize metabolites and reactions to SBRG (Schellenberger et al., 2010) abbreviations. All genome sequences were downloaded from Genbank (Benson et al., 2005) on September 21, 2012. Gene names conform to the NCBI locus name according to the original annotation in Genbank.

2.5.2 Gap Filling

The COBRA implementation of the SMILEY algorithm (growMatch) (Reed et al., 2006c) was used to predict sets of exchange and gap-filling reactions for models that were unable to simulate biomass *in silico* on M9 minimal media with glucose aerobically using FBA. The universal set of reactions used to fill gaps was the identified *E. coli* pan reactome discussed in the text. The Gurobi 5.0.0 mixed-integer linear programming solver was used (Gurobi Optimization Inc., Houston, TX) to implement SMILEY. When adding content to enable the strains to grow, exchange reactions indicating strain-specific auxotrophies were prioritized over adding new reactions without genetic evidence.

2.5.3 In silico growth simulations

Each of the 55 metabolic network reconstruction were loaded into the COBRA Toolbox (Kanehisa et al., 2010). M9 minimal media was simulated by setting a lower bound of -1000 (allowing unlimited uptake) on the exchange reactions for Ca^{2+}, Cl^{-}, CO_2, Co^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, H^{+}, H_2O, K^{+}, Mg^{2+}, Mn^{2+}, MoO_4^{2-}, Na^{+}, Ni^{2+}, SeO_4^{2-}, SeO_3^{2} and Zn^{2+}. A lower bound of -0.01 was placed on the cob(I)alamin exchange reaction. The default carbon source was glucose with a lower bound of -20, the default nitrogen source was NH_4^{-} with a lower bound of -1000, the default phosphorous source was HPO_4^{2} with a default bound of -1000 and the default sulfur source was SO_4^{2-} with a default bound of -1000. To identify sole growth supporting carbon, nitrogen, phosphorous and sulfur sources each of these default compounds were removed from the media (lower bound set to 0) one at a time and different compounds were added to determine if they supported growth. For aerobic simulations O_2 was added with a lower bound of -20 and to 0 for anaerobic simulations. For models with identified auxotrophies, the compound for which a strain was auxotrophic was also added to the M9 minimal media for each simulation with a lower bound of -10. Model growth phenotypes were determined using FBA one at a time on each condition with the core biomass reaction as the objective. Nutrient sources with growth rates above zero were classified as growth supporting, while nutrient sources with growth rates of zero were classified as non-growth supporting. The Gurobi 5.0.0 linear programming solver (Gurobi Optimization Inc., Houston, TX) was used to perform FBA.

2.5.4 Heatmap and phylogenetic tree construction

The binary results from the growth/no growth simulations for each strain were used to compute a correlation matrix based on dissimilarity indices calculated using the Jaccard method in the vegdist function of the Vegan R package. Ward's agglomerative clustering of the matrix of correlations was used to cluster the species using the hclust function of the Vegan R package and used to form a dendrogram. The heat map was visualized using the gplots R package with values aligned based on the calculated dendrogram.

2.5.5 Decision tree construction

A decision tree (**Figure 2.3**) was calculated based on growth/no growth values for each strain classified into their major pathotypes: InPec, ExPec or commensal. The classification tree tool, part of the Orange Canvas software package (Demsar et al., 2004) was used to calculate and display the decision tree using a Gini Index attribute selection criteria with no binarization and 2 minimum leaves for pre-pruning and m=2 estimate for post-pruning with leaves of the same majority class being recursively merged.

2.5.6 Strains

Eleven strains of *E. coli* and one strain of *S. flexneri* were tested for their ability to grow on different carbon sources as part of this study. The eleven *E. coli strains* are: SMS 3-5; CFT073; HS; DH1; UMN 026; K011; Sakai; ATCC 8739; 042; EDL933; and K12 MG1655. The *S. flexneri* strain was 2457T. *E. coli* 042 was a gift from Ian Henderson at Birmingham University in Birmingham, England. All other strains were purchased from ATCC.

2.5.7 Carbon source testing

Concentrated stock solutions of D-(+)-glucose, lithium acetoacetate, deoxyribose, malic acid, D-(+)-melibiose, ferric citrate, and butyric acid were made by dissolving them in M9 minimal media. Ferric citrate required heat to dissolve. The stock solutions were then filter sterilized using Millipore Millex GP 0.22 µm membranes (Millipore, Billerica, MA), after which they were diluted with sterile M9 media to a final working concentration of 20 mM. Phenylacetaldehyde and phenethylamine were dissolved directly into M9 media at a 20 mM concentration prior to filter sterilization. The M9 medium contained (per liter): 6.8 g Na_2HPO_4; 3 g KH_2PO_4; 0.5 g NaCl; 1 g NH_4Cl; 2 mM $MgSO_4$; 0.1 mM $CaCl_2$; 4.2 mg $FeCl_3$•6H2O; 45 µg $ZnSO_4$•$7H_2O$; 30 µg $CuCl_2$•$2H_2O$; 30 µg $MnSO_4$•H_2O; 45 µg $CoCl_2$•$6H_2O$; and 5.5 mg Na_2EDTA•$2H_2O$. All chemicals were sourced from Sigma-Aldrich (St. Louis, MO). Two hundred microliters of the 10 growth media were then pipetted into each row of an untreated, flat bottom 96-well plate. As a negative control, we also included one row containing M9 media only; no carbon source had been added to this row.

The eleven strains of *E. coli* and *S. flexneri* 2a strain 2457T were then tested for growth on each of the carbon sources and the negative control sample. An overnight culture of each bacterium was diluted in M9 media to an OD600 value of approximately 0.4. A 5 µL aliquot of each suspension was then inoculated into the designated wells of a 96-well plate. Growth was estimated by optical density 48 hours after inoculation. All OD600 measurements were made using a Molecular Devices Versamax plate reader. All tests were done in duplicate. Butyric acid and butane sulfonate did not support growth for any of the *E. coli* strains – including K-12 MG1655 for which the model predicted growth. This is likely because butyrate is toxic to *E. coli* cells at high concentrations such as those used in the growth screens.

2.6 Acknowledgements

We would like to thank Ian Henderson for the gift of *E. coli* strain 042. We would like to thank Kathy Andrews and Donghyuk Kim for assistance with culture and storage of the strains used in this work. Additionally, we would like to thank Aarash Bordbar for helpful discussions and insights. This work was funded by grants 1R01GM098105 and 1R01GM057089 from NIH/NIGMS.

Chapter 2, in full, is a reprint of the material Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BO: Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A* 2013, 110(50):20338-20343. The dissertation author was the primary author.

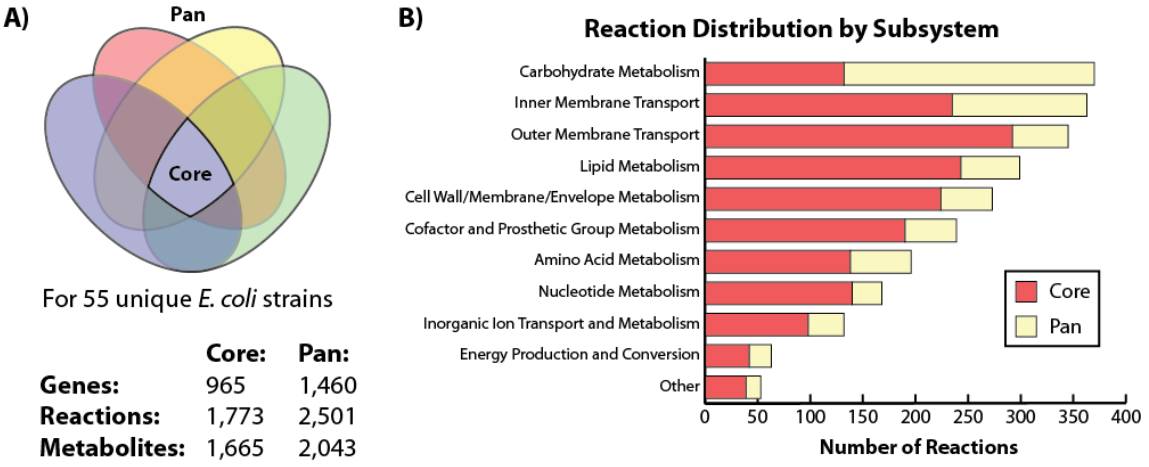


Figure 2.1. Core and Pan metabolic capabilities of the *E. coli* species
The core and pan metabolic content was determined for genome-scale metabolic models (GEMs) of 55 unique *E. coli* strains. **A)** The core content, illustrated by the intersection of the Venn diagram, is shared with all strains. The pan content consists of all content in any model and includes the core content. The Venn diagram is not to scale. **B)** Classification of reactions in the core and pan reactomes by metabolic subsystem.

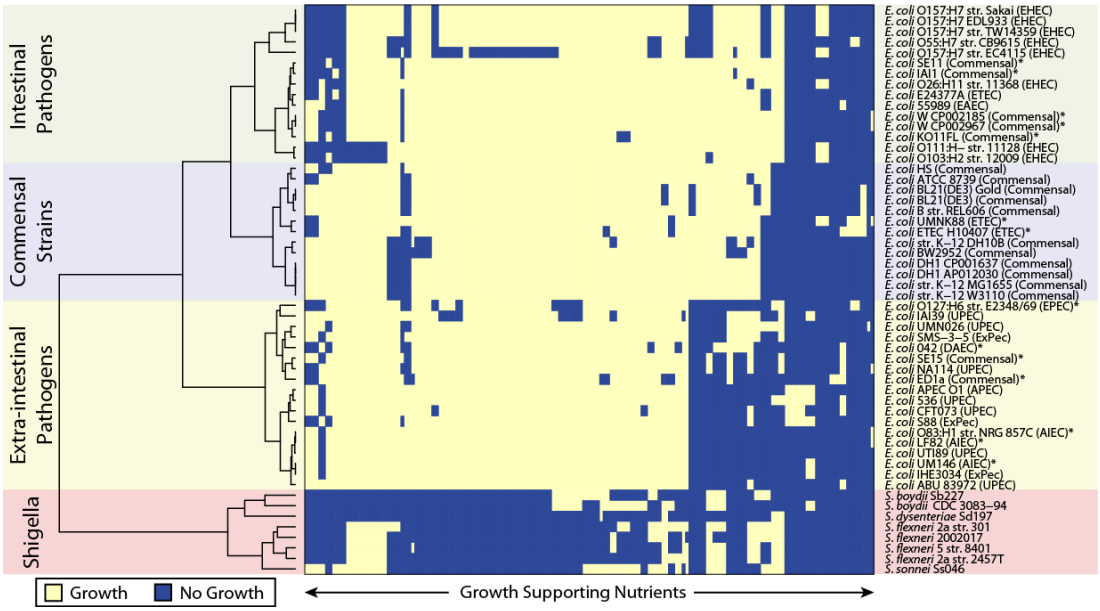


Figure 2.2. Clustering of species by unique growth-supporting conditions.
Predicted metabolic phenotypes on the variable growth-supporting nutrient conditions composed of different carbon, nitrogen, phosphorous and sulfur nutrient sources in aerobic and anaerobic conditions. Strains are clustered based on their ability to sustain growth in each different environment. Rows represent individual strains and columns represent different nutrient conditions. In general, strains clustered into their respective pathotypes of commensal *E. coli* strains, intestinal pathogenic *E. coli* strains, extra-intestinal pathogenic *E. coli* strains and Shigella strains. An asterix symbol indicates those strains that clustered outside of their respective pathotype.

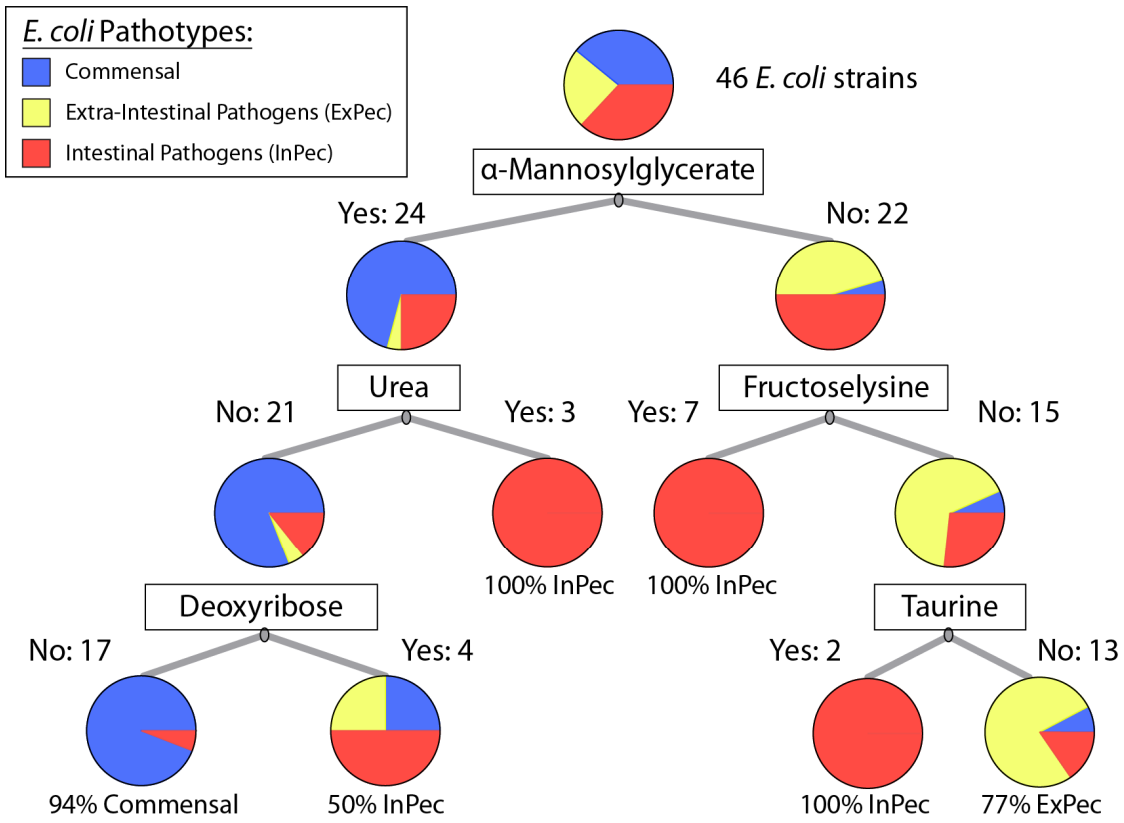


Figure 2.3. Classification of *E. coli* pathotypes based on growth-supporting conditions
Growth-supporting nutrients were used to create a classification tree. This tree can be used to determine if an *E. coli* strain is commensal, an intestinal pathogen or an extra-intestinal pathogen. For example, following the tree to the right shows that 77% of *E. coli* strains that cannot grow on α-mannosylglycerate, fructoselysine, or taurine as sole carbon sources are expected to be Extra-intestinal pathogens. Thus, a small number of nutrient sources can be used to classify *E. coli* strains of different types.

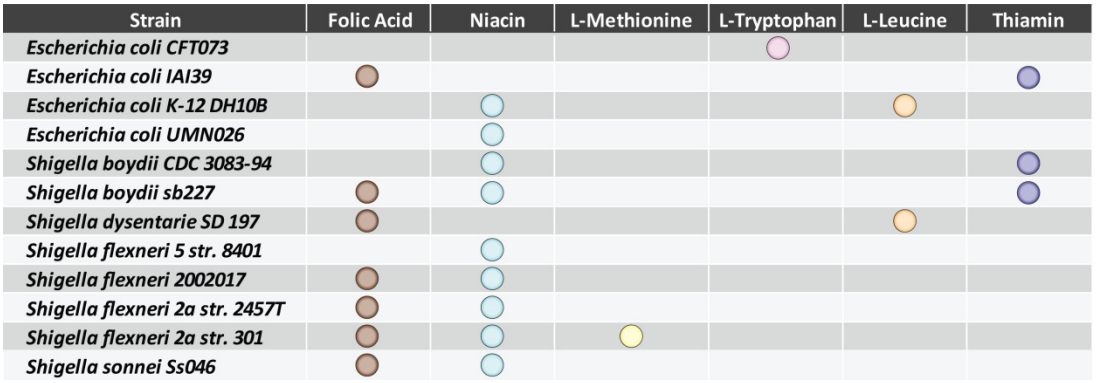
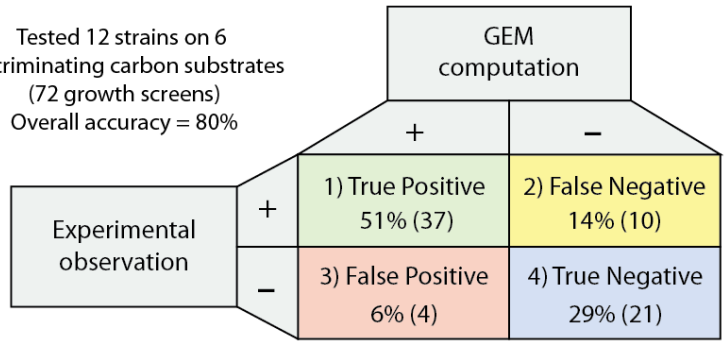


Figure 2.4. Model predicted strain-specific auxotrophies
GEM predicted minimal media conditions for each auxotrophic strain. *Shigella* strains lack essential vitamin biosynthesis capabilities for niacin (vitamin B3), thiamin (vitamin B1) and folate (vitamin B9). Other strains have lost biosynthetic pathways for the essential amino acids methionine, tryptophan and leucine, thus becoming auxotrophic for these compounds.

A)



B)

		Commensal Strains					Intestinal Pathogens			Extra-intestinal Pathogens			Shigella	
Compound	**Genes**	MG1655	8739	DH1	HS	KO11FL	O42	EDL933	Sakai	SMS3-5	CFT073	UMN026	2457T	**Substrate accuracy**
Acetoacetate	*atoE, atoAD*	TP	TP	TP	TP	TN	TP	TN	TN	TP	TP	TP	TN	100%
Deoxyribose	*deoC, deoP, deoK*	TN	TN	TN	TN	TN	TP	TN	TN	TN	TP	TN	TN	100%
D-malate	*dctA, yeaU*	TP	TP	TP	FP	TP	TP	TP	TP	TP	FP	TP	TN	83%
Ferric dicitrate	*fecBCDE*	TP	TP	TP	TN	TN	TP	TN	TN	FN	TN	FP	TN	83%
Melibiose	*melA, melB*	TP	TP	FP	TP	TP	FN	TP	TP	TP	FN	TP	TP	75%
2x Phenyls	*tynA, feaB, paaK*	TP	TP	TP	TP	TP	FN	FN	FN	FN	FN	FN	FN	42%
	Strain accuracy:	100%	100%	83%	83%	100%	67%	83%	83%	67%	50%	67%	83%	

Figure 2.5. Comparison of GEM predictions to experimental results
A) Comparison of GEM predictions to experimental results revealed a high level of accuracy (80%) both true positives (Quadrant 1) and true negatives (Quadrant 4). False negative cases (Quadrant 2) represent missing knowledge and are an opportunity for biological discovery. False positive cases (Quadrant 3) represent missing context-specific information such as transcriptional regulation. **B)** A detailed breakdown of the comparisons based on the pathways (rows) and screened strains (columns).

Table 2.1. GEM predicted advantages for sole growth-supporting nitrogen and carbon sources between commensal and extra-intestinal pathogenic *E. coli* (ExPec) strains

Source	Commensal	ExPec	P-value
A. ExPec strain nutrients:			
3-Phospho-D-glycerate	0%	36%	1.3E-2
L-Arginine exchange	11%	64%	5.0E-3
Cellobiose exchange	33%	82%	1.3E-2
N-Acetyl-D-galactosamine	67%	100%	3.9E-2
B. Commensal strain nutrients:			
Fructoselysine	89%	0%	2.2E-6
Psiscoselysine	89%	0%	2.2E-6
Dopamine	89%	0%	2.2E-6
Phenethylamine	89%	0%	2.2E-6
Tyramine	89%	0%	2.2E-6
Phenylacetaldehyde	72%	0%	1.2E-4
Alpha-Mannosylglycerate	94%	9%	5.7E-6
4-hydroxyphenylacetate	56%	9%	1.3E-2
Cyanate	83%	27%	3.8E-3
Melibiose	78%	27%	9.7E-3
Phenylpropanoate	72%	27%	2.0E-2
3-(3-hydroxy-phenyl)propionate	89%	36%	5.0E-3
3-hydroxycinnamic acid	89%	36%	5.0E-3

Table 2.2. GEM predicted advantages for sole growth-supporting nitrogen and carbon sources between commensal and intestinal pathogenic *E. coli* (InPec) strains.

Source	Commensal	InPec	P-value
B. InPec strain nutrients:			
Sucrose	33%	65%	5.0E-2
Raffinose	28%	65%	2.6E-2
L-Arginine	11%	65%	1.3E-3
D-Arabitol	0%	24%	4.6E-2
Ribitol	0%	24%	4.5E-2
Agmatine	0%	47%	1.0E-3
Urea	0%	47%	1.0E-3
A. Commensal strain nutrients:			
Galactonate	100%	65%	7.6E-3
α-Mannosylglycerate	94%	35%	2.6E-4
Dopamine	89%	41%	3.5E-3
Phenethylamine	89%	41%	3.5E-3
Tyramine	89%	41%	3.5E-3
5-Dehydro-D-gluconate	78%	24%	1.6E-3
L-Idonate	78%	24%	1.6E-3
D-Allose	78%	41%	2.5E-2
Butyrate	78%	47%	5.0E-2
Acetoacetate	78%	47%	5.0E-2
4-hydroxyphenylacetate	56%	18%	2.0E-2

Chapter 3 Comparative genome-scale modelling of multiple *S. aureus* strains identifies strain-specific pathogenic characteristics and unique metabolic capabilities

3.1 Abstract

Staphylococcus aureus is a preeminent bacterial pathogen capable of colonizing diverse ecological niches within its human host, including the respiratory tract, skin and nasal passages. Possessing numerous immune resistance factors and toxins, *S. aureus* is a leading cause of skin and soft tissue infections, pneumonia, sepsis and endocarditis. Numerous studies have focused on *S. aureus* genomics, molecular epidemiology and antibiotic resistance; however, few studies have analyzed the totality of the pathogen's metabolic functions and virulence capabilities at the genome-scale. To fill this gap, we constructed genome-scale models (GEMS) of 64 diverse *S. aureus* strains that spanned ecological niches, host types and antibiotic resistance profiles. The GEMs enabled a systems approach to characterizing the pan and core metabolic capabilities of the *S. aureus* species. We found that *S. aureus* has an open pan-genome comprised of 7,411 genes for a chosen set of 64 *S. aureus* strains and a core-genome composed of 1,441 genes. GEMs were used to systematically predict growth capabilities in more than 300 different growth-supporting environments. Most *S. aureus* strains show similar metabolic capabilities, but some important differences were identified in amino acid and nucleotide biosynthesis pathways, including predicted strain-specific auxotrophies. All models were predicted to be auxotrophic for the vitamins niacin (vitamin B3), thiamin (vitamin B1) and riboflavin (vitamin B2). GEM computations led to the identification of 148 core essential genes. In addition, a virulome of 81 known *S. aureus* virulence factors was constructed and compared across the strains. Of these 81 genes, 27 were found in all strains, forming a core virulome. The core and pan metabolic capabilities, combined with presence or absence of specific virulence factors, can be used to classify *S. aureus* strains according to their preferred host and infectious niche.

3.2 Background

Staphylococcus aureus is a Gram-positive pathogen responsible for several infection types in humans, ranging from simple skin infections, such as boils, to more complicated ones like pneumonia, endocarditis and necrotizing fasciitis (Davis et al., 2007). *S. aureus* can reside asymptomatically in the human nostril, and it has been estimated that approximately 30% of humans are healthy (Kluytmans et al., 1997). In recent years, clinical management of this leading pathogen has been complicated by its continuous acquisition of resistance to front-line antibiotics. Resistance to all antibiotic classes has developed rapidly in sequential epidemic waves (Chambers and Deleo, 2009), starting from the mid-1940s, with the emergence of penicillin resistance to the late 1990s, characterized by the emergence and wide dissemination of community-associated methicillin-resistant *S. aureus* (CA-MRSA) and continued sporadic reports of vancomycin resistance in hospital settings. In the last decade, MRSA alone has caused more than double the number of invasive infections than other major pathogens including *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis (Klevens et al., 2007)*. The emergence of CA-MRSA represented a particularly worrisome global threat because serious infections can occur in healthy individuals (Herold et al., 1998), highlighting their enhanced virulence compared to many forms of hospital-associated MRSA (HA-MRSA). Different clones of CA-MRSA have been identified in the United States, Canada, Asia, South America, Australia and throughout Europe, including countries with historically low prevalence of MRSA, such as Norway, Netherlands, Denmark and Finland (Larsen et al., 2008, Laupland et al., 2008, Wannet et al., 2005) . Global spread of MRSA infections led to increased vancomycin usage against serious infections, promoting the emergence of vancomycin-

intermediate (VISA) and resistant (VRSA) strains (Fridkin et al., 2003, Hiramatsu et al., 1997). For these reasons, *S. aureus* remains one of the most challenging pathogens of our time.

Despite many studies focusing on S. *aureus* genomics, molecular epidemiology and mechanisms of drug resistance and cytotoxicity, there are relatively few studies that examine *S. aureus* basic biochemistry and metabolic function on a genome-scale. One way to accomplish this is the use of Genome-scale Network Reconstruction (GENRE) (Becker and Palsson, 2005). GENREs represent all of the biochemical reactions occurring in an organism directly connected to their genetic basis. GENRE can be expressed mathematically in the form of a genome-scale model (GEM) which can be computationally analyzed by various methods (Lewis et al., 2012a, Price et al., 2004) that predict experimentally testable phenotypic properties (Bordbar et al., 2014).These predictions include growth capabilities on different nutrient sources, allowing identification of strain-specific auxotrophies, and prediction of essential genes and reactions required to sustain *in silico* microbial growth. Such analyses have led to potential targets for enzymatic inhibition using ad hoc designed drugs (Becker and Palsson, 2005, Lee et al., 2009, Schmidt et al., 2013b, Aziz et al., 2015a). Considering the global health threat posed by drug-resistant strains (WHO, 2014), identification of putative weak points in *S. aureus* metabolism would represent a major step forward in control of MRSA infections.

Becker et al. provided a well-curated GENRE of *S. aureus* N315 (named iSB619) that represented the first biochemically, genomically and genetically structured knowledge base for *S. aureus* metabolism. However, knowledge from one strain is never sufficient to represent an entire species (Monk et al., 2013). Currently there are 49

complete genome sequences available for *S. aureus* strains in NCBI Genome database (Benson et al., 2009) with an additional 450 draft sequences for comparison to provide comprehensive insight into the pan-genome of the *S. aureus* species. A bacterial species can be effectively described by its pan-genome (Tettelin et al., 2008), which can be divided into the core genome (genes shared by genomes of all strains in the species and that are likely to encode functions related to basic cellular biology) and the dispensable genome (genes present in some, but not all, the representatives of a species (Medini et al., 2005). The dispensable genome likely includes functions that confer specific advantages under particular environmental conditions, such as adaptation to distinct niches, antibiotic resistance, and the ability to colonize new hosts or proteins/functions that are recognizable by phages or immune cells, and are thus under positive selective pressure.

In this study we used three different GEMs available for *S. aureus* N315 (Becker and Palsson, 2005, Heinemann et al., 2005, Lee et al., 2009) to build a reference GEM that served as a platform for development of GEMs for an additional 64 *S. aureus* strains. Strains were selected to provide a heterogeneous dataset of the species. We expanded and updated the reference GEM based on new annotations, recent literature and metabolic databases, thus providing a reliable foundation for the development of the other strain-specific GEMs. Together, the genomic sequences and the GEMs provide two different and complementary ways to explore diversity within the *S. aureus* species and to compare core vs. pan genomic functionality and metabolic capabilities. Furthermore, we used the GEMs to simulate phenotypes of different *S. aureus* strains in a variety of environments and to compute essential genes and biochemical reactions. The collection of *S. aureus* GEMs were also used to develop a core GEM representing

the conserved metabolic capabilities of the *S. aureus* species, which, in principle, can be used to design targeted therapeutics against the pathogen (Kim et al., 2010a).

3.3 Results

3.3.1 Building an initial reconstruction of *S. aureus* as a species

A set of 225 publically available *S. aureus* genome sequences were downloaded from the NCBI including all 48 completely assembled genomes (Geer et al., 2010). From this set, a phylogenetic tree was constructed using the concatenated sequence of 7 conserved housekeeping genes (*arcC, aroE, glpF, gmk, pta, tpi* and *yqiL*) commonly used to define clonal complexes in clinical studies of *S. aureus* also known as multi-locus sequence tags (MLST) (**Figure 3.1A**). The most distantly related strain was the Australian *S. aureus* isolate MSHR1132 belonging to the clonal complex 75 lineage. The average nucleotide divergence between orthologous genes in MSHR1132 and typical *S. aureus* is approximately sevenfold greater than the maximum divergence previously observed (Holt et al., 2011). Next, a representative set of 64 *S. aureus* strains were filtered from this larger dataset for further analysis. Strains were selected based on four criteria: i) drug resistance (MRSA**:** methicillin-resistant *S. aureus***,** MSSA**:** methicillin-sensitive *S. aureus***,** VRSA**:** vancomycin-resistant *S. aureus***,** VISA**:** vancomycin-intermediate *S. aureus***)** (**Figure 3.1B**), ii) host specificity (human vs. animal) (**Figure 3.1C**), iii) epidemiological source (CA-MRSA: community-associated MRSA, HA-MRSA: healthcare-acquired MRSA, LA-MRSA: livestock-associated MRSA) (**Figure 3.1D**), and iv) evolutionary distance based on tree topology. Thus, the topology of the tree coupled with clinical and epidemiological phenotypic information allowed for development of a heterogeneous dataset spanning representative strains of the *S. aureus* species.

3.3.2 Characteristics of the *S. aureus* core and pan-genomes

To gain insight into the genomic diversity of the selected dataset of 64 *S. aureus* strains, we used them to compute the pan-genome of the species. Gene content of the pan-genome can be used to effectively describe a bacterial species. The pan-genome is divided into three sections: i) the core-genome (the set of genes shared by virtually all strains in a species), ii) the accessory genome (the set of genes present in some, but not all, representatives of a species) and iii) the unique genome (genes unique to individual members of the species, commonly known as ORFans or singletons). Because the pan-genome is made up of thousands of genes, it offers a much higher resolution for strain typing compared to genotyping techniques such as MLST (Hall et al., 2010).

The 64 *S. aureus* strains examined yielded a pan-genome size of 7,457 genes. This set can be divided into the core, accessory and unique genomes. Based on this dataset, the core-genome is composed of 1,441 genes, the accessory genome is composed of 2,871 genes and the unique genome is composed of 3,145 genes (**Figure 3.2A**). To estimate the core-genome size, the asymptotic limit of the pan-genome size and the expected rate of discovery of novel genes, we used a sampling approach to obtain a comprehensive set of randomly permuted pan-genomes, measuring the number of total genes, shared genes and new genes, respectively. By fitting a double exponential decay function to the number of shared genes, we estimate the *S. aureus* core genome to have 1,425 genes (**Figure 3.3**). New genes and total genes were similarly used to obtain two additional fits of the Heap's law function, resulting in γ values of -0.58 and 0.31, respectively. The γ parameter determines the behavior of the curve. For γ values > 0, the function does not have any asymptote indicating that the *S. aureus* gene repertoire is likely to grow indefinitely as more strains are sequenced. Indeed, the

best fit for total genes was obtained with γ value = 0.23, which denotes an open pan-genome, while for the novel genes the best fit was obtained with γ value = 0.58.

Functional annotation of genes in the pan-genome was achieved using the COG database (Tatusov et al., 2003) and is shown in **Figure 3.2**, revealing a heterogeneous distribution of functional categories among the three different pan-genome sets. When genes were classified into metabolic and non-metabolic, the largest fraction of the core genome consists of metabolic genes (58%), this consisted of primarily genes involved in central metabolism including glycolysis/gluconeogenesis, the pentose phosphate pathway and TCA cycle. A much lower fraction of metabolic gene content is observed in the accessory and unique genome (33% and 18%, respectively).

The average *S. aureus* genome encodes 2,800 genes; therefore the size of the core-genome represents a high portion (56%, on average) of each *S. aureus* genome (**Figure 3.2B**). The fact that each *S. aureus* strain has such a high portion of shared genes can be interpreted as a direct consequence of the proposed clonal structure for the species (Feil et al., 2003) . The majority of 1,441 core genes (shared by all 64 *S. aureus* strains examined) are involved in housekeeping processes. These include non-metabolic functions such as transcription (15%), translation, ribosomal structure and biogenesis (14%), RNA processing and modification (7%), and genes associated with metabolic functions such as amino acid transport and metabolism (11%), carbohydrate transport and metabolism (7%), coenzyme transport and metabolism (4%), cell wall and membrane biosynthesis (5%) and energy production and conversion (12%).

Completely different gene functional assignments were observed for those genes in the accessory and unique genomes, with the accessory genome comprising mostly genes associated with mobile genetic elements such as transposons and

bacteriophages (replication, recombination and repair 24%) or with non-metabolic functions such as transcription (14%) and translation (10%). Metabolic functions were also included in the accessory genome, including those related to amino acid metabolism (8%), inorganic ion transport (7%) and carbohydrate metabolism (6%). The unique genome is heavily enriched in genes related to mobile elements (62%), with the other categories being poorly represented. This high proportion of mobile elements in the unique genome is similar to other organisms, including *E. coli*, where horizontal gene transfer has had a large impact on species evolution (citation).

3.3.3 Analysis of atypical *S. aureus* genes

To investigate the impact of HGT events on the *S. aureus* species, we searched for atypical genes within each strain by computing DNA composition (GC content) differences between core and pan genes as well as by using a Hidden Markov Model (HMM) based tool (SIGI-HMM) (Waack et al., 2006) that detects genomic islands based on statistical analysis of codon usage with high precision (Langille et al., 2008). A total of 4,277 atypical genes were identified based on GC and the average genome had 49 unique genes present **Figure 3.2B**). Based on the codon analysis we identified 1,788 unique genes with an average of 28 unique genes per genome.

We assigned functional categories to these atypical genes and examined the distribution of genes assigned to each functional category in the *S. aureus* genomes. Most of the identified atypical genes have no known COG functional class. The phylogenetic origin of each atypical *S. aureus* gene was traced using the nr-database. For each gene, the best non-*S. aureus* hit was taken as the *putative transfer source* (PTS). We hypothesized that more recent HGT events were indicated by genes with fewer *S. aureus* hits, i.e. those more similar to the query gene than the PTS. We used

this number as an index to estimate the *ancestrality* of the HGT events for each PTS, a measurement we term the Ancestrality Index (AI). It was only possible to find a significant PTS for 25% of the atypical genes we identified. Those PTS with an AI > 2 mostly corresponded to HGT events that occurred at the genus level, while the majority of the more "recent" PTS (an AI <= 2) corresponded to HGT events at higher taxonomical levels (order, class and phylum). In this group we found a high representation of proteins such as hypothetical proteins, mobilization proteins, metal and antibiotic resistance genes and virulence factors.

3.3.4 Characteristics of *S. aureus* core and pan metabolic content

Because so much of the *S. aureus* core-genome is dedicated to metabolic functions, metabolic characteristics of the sequenced *S. aureus* strains were studied to further develop a definition of a *S. aureus* species and its metabolic capabilities. The 64 *S. aureus* genome sequences were used to construct strain-specific genome-scale metabolic reconstructions that were used to compare gene, reaction, and metabolite content between the strains. Each reconstruction serves as a comprehensive representation of the metabolic capabilities of an *S. aureus* strain. Content shared among all reconstructions defines the 'core' metabolic capabilities of the *S. aureus* species. Similarly, the metabolic capabilities of all strains were combined to define the full potential of metabolic capabilities for the *S. aureus* species, or its 'pan' metabolic network (**Figure 3.3A**) (Monk et al., 2013).

The size and content of the core metabolic network characterizes the metabolic foundation of *S. aureus* as a species, comprising 729 metabolic genes that catalyze 1,411 reactions involving 1,232 metabolites. Highly conserved metabolic subsystems across all *S. aureus* models included lipid metabolism, energy metabolism, glycan

biosynthesis and metabolism of polyketides and terpenoids. Reactions involved in these metabolic subsystems were highly represented (>95% conserved) in the core "reactome". By contrast, only 80% of amino acid biosynthesis reactions were part of the core reactome. Conserved amino acid biosynthesis pathways included those for valine, leucine and isoleucine, and alanine (but not aspartate) biosynthesis. The core reactome also contained catalase, an enzyme that hydrolyzes hydrogen peroxide into water and oxygen and that is used in the clinical laboratory to distinguish staphylococci from enterococci and streptococci. Catalase production and oxidant resistance have been shown to be a predisposing factor for nasal colonization and subsequent infection (Park et al., 2008).

The pan metabolic capabilities are comprised of the total number of different reactions found in all strains and thus indicate the pool of metabolic capabilities within a species. The *S. aureus* pan metabolic network contains 877 metabolic genes, 1,524 reactions and 1,313 metabolites. About 20% of reactions in amino acid metabolism were specific to the pan reactome, forming the largest group (**Figure 3.3B**). A majority of these reactions are involved in arginine and proline metabolism, histidine metabolism and tryptophan metabolism. Furthermore, metabolic reactions for the synthesis of amino acids L-proline, L-cysteine and L-Leucine were lacking from either the core or pan reactome, indicating that *S. aureus* as a species is incapable of synthesizing these amino acids in agreement with previous studies (Lee et al., 2009).

3.3.5 Determining strain-specific auxotrophies

The conversion of metabolic network reconstructions into computable mathematical models enables computation of phenotypes in diverse nutrient environments. The 64 *S. aureus* strain-specific reconstructed networks were converted

into genome-scale metabolic models (GEMs) that were used to interpret the content of each reconstruction and predict strain-specific nutritional requirements. The GEMs use a biomass objective function of to represent cell growth. It consists of 58 metabolites including amino acids, lipids, nucleotides and cell wall components that all must be produced for cellular growth that have been measured in specific ratios for various *S. aureus* strains (Becker and Palsson, 2005, Heinemann et al., 2005).

Reactions belonging to the amino acid metabolism subsystem made up the majority of reactions in the pan-reactome (**Figure 3.3**). We hypothesized that functional differences in amino acid biosynthesis capabilities of different strains of *S. aureus* may allow different strains to adapt to different nutritional environments. To test this hypothesis, we simulated growth *in silico* for all 64 *S. aureus* GEMs on a variety of minimal media growth conditions, including a minimal growth media reported for *S. aureus* N315 (Becker and Palsson, 2005).

Computing *in silico* growth rates can identify strain-specific auxotrophies. These auxotrophies arise when a model lacks the metabolic capabilities to synthesize biomass components from the minimal media components or when a gap in the metabolic network disrupts a metabolic pathway central for the production of biomass components. The *in silico* growth analysis revealed that all 64 models i) require a minimal set of nutrients including vitamin B1 (thiamin), vitamin B2 (riboflavin), vitamin B3 (niacin) and spermidine ii) require at least one of four amino acids (proline, arginine, glutamine or ornithine) for growth, iii) were auxotrophic for methionine and cysteine. Adding these amino acids to the *in silico* media alone was not sufficient to replace those listed previously.

There were also several strain-specific auxotrophies, with 23 of 64 models unable to grow with the minimal media defined. These models lacked the ability to synthesize additional compounds including uracil, putrescine and the amino acids arginine, histidine and tryptophan (**Figure 3.4**). Moreover, 8 closely related strains, members of the same clade, also known as the Italian or South German clone (citation), were all auxotrophs for tryptophan. We tested these predicted auxotrophies in four strains of *S. aureus* (USA300, N315, Mu50 and 8325-4) and found that none of these strains grew in a minimal chemically defined media supplemented with proline, serine and leucine. However, when threonine was added to the media, N315 and USA300 were able to grow. Addition of arginine, phenylalanine or valine to the minimal media did not support growth of any of the strains, but addition of all 20 amino acids supported growth in all four strains (**Supplementary Figure 1**).

3.3.6 Calculating alternative nutrient sources

The 64 *S. aureus* GEMs were used to predict growth capabilities on alternative carbon, nitrogen, phosphorous and sulfur sources by removing glucose, ammonia, sulfate and phosphate from the *in silico* growth media and adding alternative sources one at a time. A total of 300 alternative nutrient sources were tested using flux balance analysis (FBA) (Orth et al., 2010b)to assess whether each S. *aureus* strain grew *in silico*, allowing grouping of strains according to predicted growth abilities. Here very low variability was identified in terms of carbon usage, reflecting minimal differences among the metabolic capabilities of the strains analyzed. We specifically looked to see if the models were predicted to be able to utilize arginine for growth. The arginine mobile catabolic element (ACME) is present in MRSA USA300 strains (citation), however we

found that the argine catabolic capability, ecoded for by ArcA-ArcD was conserved across all *S. aureus* strains consistent with previous work (Zhu et al., 2007).

3.3.7 Prediction of essential metabolic genes

The effects of gene deletion in a metabolic network can be determined computationally. In this way, one can predict a set of genes essential for the production of biomass and growth of the organism. A gene is considered essential if its deletion *in silico* blocks production of biomass by the GEM. Such genes represent good targets for *ad hoc* design of enzymatic inhibitors. Additionally, essential genes predicted for each strains can be used to identify: i) the set of essential genes shared between all the strains ("core essential genes"), which could be used as a testing pool for designing drugs that target all *S. aureus* representatives, and ii) the set of essential genes specific to some strains ("unique essential genes"), which can be used to design strain-specific drugs. These computations identified a total of 148 core essential genes.

3.3.8 S. aureus strains share a core set of ubiquitous genes encoding proteins involved in transcription and translation

Most modern antibiotics do not only target an organism's metabolic functions, but also target the transcriptional and translational machinery of the target organism. We compared the conservation of transcriptional and translational machinery within *S. aureus* strains across the species. Genes involved in these processes were curated from *E. coli* and *B. subtilis* because *E. coli* is the organism for which almost all components of transcription and translation machinery have been identified and experimentally characterized. However, because *S. aureus* is phylogenetically closer to *B. subtilis* (both are Firmicutes), additional proteins from *B. subtilis* were used to assemble this dataset. Although *B. subtilis* homologs exist for most of the *E. coli* genes

involved in transcription and translation, a few *B. subtilis* genes exist for which no homologs are found in the *E. coli* genome and vice versa. Altogether we selected 289 query genes of which 192 are shared between *E. coli* and *B. subtilis* while 59 are unique to *E. coli* and 42 are unique to *B. subtilis*. The set included genes that encode functions for transcription, ribosome biogenesis, tRNA maturation and aminoacylation, and proteins and cofactors required for mRNA translation and RNA decay. Using these experimentally validated genes as input (Grosjean et al., 2014), genes encoding proteins of the core transcription and translation machinery were predicted in the 64 *S. aureus* strains.

A core set of 239 genes involved in transcription and translation was found in all *S. aureus* strains examined. The majority of genes coding for ribosomal proteins, aminoacyl-tRNA synthetases, translation factors and several ribosome biogenesis/maturation enzymes are universally conserved in *S. aureus* strains. These same genes are essential in both *E. coli* and *B. subtilis* (Fang et al., 2005, Koonin, 2003, Mushegian, 2008) and other bacterial organisms (Ciccarelli et al., 2006, Yutin et al., 2012). Conversely, 47 of the 289 genes were absent in all *S. aureus* strains examined. Of genes absent in all *S. aureus* strains examined, seven are present in both *E. coli* and *B. subtilis, 33* are unique to *E. coli* and seven are unique to *B. subtilis*. Most of these missing genes encode functions in transcription, tRNA modifications, rRNA modifications and RNA processing.

Next we examined genes present in some *S. aureus* strains, but not others. These included the ribonuclease RNE involved in RNA processing and the 50s ribosomal protein L33 (RPMGA). These two proteins were conserved in 28 of the *S. aureus* strains examined. RMPGA was missing in 19 strains and RNE was missing in

10 strains. While most of the *S. aureus* strains retained the two genes encoding L33a (RpmGa) and L33b (RpmGb), L33a was lost in 19 of the 64 strains examined. L33 is responsible for cellular ribosome heterogeneity and may generate specialized ribosomes in response to stress conditions and environmental changes (Byrgazov et al., 2013). These proteins could have evolved to fulfill specific non-essential innovation (Lecompte et al., 2002) and hence easily be lost in reductive evolutions. The L33 gene is also non-essential in *E. coli* and *B. subtilis* (Akanuma et al., 2012, Baba et al., 2006, Bubunenko et al., 2007, Shoji et al., 2011). Our analysis defines the minimal and conserved set of genes needed to encode functions that sustain protein synthesis in various *S. aureus* strains. Because the constraints-based reconstruction and analysis approach has been applied to macromolecular synthesis via the processes of transcription and translation (O'Brien et al., 2013b, Liu et al., 2014a) genes identified in this study serve as a starting point for reconstructed of an expression matrix 'E' matrix for *S. aureus*. A detailed discussion of these 'E' genes present in *S. aureus* is provided in the supplementary text. Therefore inhibitors of these proteins may be good targets for future antibiotic therapies against *S. aureus*.

3.3.9 Construction of a virulome for the *S. aureus* species

To gain insight into the conservation of virulence factors across the *S. aureus* species we curated a set of virulence factors (VFs) present in different strains based on literature and database searches (Chen et al., 2012b). This set of VFs forms a comprehensive "virulome" of the *S. aureus* species. The virulome comprised a total of 90 different VFs that are uniquely present in the different *S. aureus* strains examined (**Figure 3.5**).

Of the 90 VFs, 35 were shared by all of the strains forming a core set of VFs. Nine of the conserved VFs are cap8 genes (B, C, E, F, L, M, N, O and P), which are involved in

the synthesis of the polysaccharide capsule (PC). The PC evolved in *S. aureus* to impair the opsonophagocytic uptake by neutrophils, and the vast majority of clinical isolates are known to express surface capsules composed of serotype 5 or 8 polysaccharide (Nanra et al., 2013). Because the PC genes are found in the core virulome, all the *S. aureus* strains represented in our dataset are potentially able to synthesize either the cap5 or cap8 PC.

Other conserved VFs included five involved in the production of two different cytotoxins, namely the Panton-Valentine leukocidin (PVL), encoded by the genes *lukS* and *lukF,* and the gamma-hemolysin, encoded by *hlgA*, *hlgB* and *hlgC*. Six other conserved VFs encode iron-regulated proteins (*isdA*, *isdC*, *isdE* and *isdF*) that bind to extracellular matrix components such as fibrinogen and fibronectin to promote cell adherence. Among these, the *isdA* gene plays a role in the *S. aureus* iron acquisiton system, important for *S. aureus* in vivo replication and disease pathogenesis (Kehl-Fie and Skaar, 2010). In vertebrates, the vast majority of iron is stored within the cells, but bacterial cytotoxins can damage host cells including erythrocytes to liberate iron. The host releases lactoferrin during the innate immune response to restrict iron availability and provide antimicrobial activity, but IsdA can bind to lactoferrin conferring resistance to this host defense pathway (Clarke and Foster, 2008) .

Additional conserved virulence factors include the *ica* genes (*icaA* and *icaR*) involved in the production of the polysaccharide poly-***N***-acetylglucosamine (PNAG), which promotes biofilm formation (Cerca et al., 2008). Additional core virulence determinants included *hysA* encoding hyaluronate lyase, which cleaves the extracellular matrix glycosaminoglycan within the connective tissues, and *sbi* encoding an IgG binding protein that thwarts immunoglobulin host defense. Other canonical *S. aureus* virulence

factors were highly conserved but were not found across all 64 strains. Protein A (binds immunoglobin G to disrupt phagocytosis, encoded for by *spa*) was found in 90% of strains. Alpha toxin (disrupts the membrane and enhanced invasiveness, encoded for by *hla*) was found in 96% of strains. Biosynthesis genes for the staphyloxanthin pigment discussed above (encoded for by *crtMNPQ*) were found in 95% of strains. A previous analysis of clinical *S. aureus* isolates found that 90% were positive for staphyloxanthin, agreeing with the results presented here.

Several other VFs were strain-specific. For example, the SCIN protein (Staphylococcal complement inhibitor, encoded by *scn*) that was present in 48 of the 64 strains and CHIPS (Chemotaxis inhibitory protein of staphylococci, encoded by *chp*) was present in 27 of the 64 strains. The AGR quorum sensing system (regulates biofilm development, encoded for by *agrABCD*) was found in 25% of strains Meanwhile, the exfoliative toxin B was found only in in a single strain (11819-97). *S. aureus* MW2 was found to have the most established virulence factors (79 VFs) followed by MSSA476 (74 VFs) and Mu3 and Mu50 (70 VFs). In contrast, the ST288 strains had the fewest number of VFs (53-54 total). These VFs can be used to classify strains of *S. aureus* based on lifestyle and niche. For example, most livestock associated strains of *S. aureus* can be distinguished from human associated strains by searching for the presence only three VFs, staphylokinase precursor (*sak*), staphylococcal enterotoxin (*seg2*), Immunoglobulin G binding protein A precursor (*spa*) (**Figure 3.5B**).

3.4 Discussion

In this study, we compared the genomes of 64 strains of *S. aureus* to gain insights into the diversity of this species from different perspectives. We used both genomic comparisons and metabolic modeling to gain insight into the metabolic and

genetic diversity that define different strains of the species. Using a comparative genomics approach, we computed a pan-genome of the *S. aureus* species, which, to the best of our knowledge, provides the most comprehensive pan-genome produced for this taxon. Identification and characterization of the makeup of a species' pan-genome is a powerful tool to analyze genomic diversity within a clade, and can be used to predict, via extrapolation, the number of genome sequences required for bounding the gene repertoire of any analyzed clade. Our regression analysis shows that the *S. aureus* pan-genome is open, indicating that the gene repertoire of this species is theoretically boundless. This result is in agreement with a previous DNA microarray experiment involving 36 *S. aureus* strains (Fitzgerald et al., 2001), in which extensive genetic variability was reported.

The pan and core-genome comprise respectively 7,457 and 1,441 clusters of orthologous genes, which were functionally classified according to the COG database. The different functional distributions in the pan-genome categories (core, accessory and unique) revealed that core genes are mostly associated with housekeeping functions (i.e. control of gene expression machinery and basic biochemistry), while accessory and unique genes are more associated with niche specific functions including degradation of antibiotics and production of different antimicrobial compounds. In particular, a high portion of unique genes were found to be related to mobile genetic elements (i.e. transposon, phages, plasmids), that may drive acquisition of novel functional modules via HGT, including drug resistance and virulence (McCarthy and Lindsay, 2012). Beyond these evolutionary insights, the pan-genome has important practical implications. For instance, the pan-genome can be used to design vaccines by reverse vaccinology (Donati and Rappuoli, 2013). This approach has been successful for influenza and

streptococcus where polyvalent vaccines have been designed to target specific strains of the species based on a search of the pan-genome to identify accessory genes that are only present to those strains that are most dangerous. Such efforts are critical for *S. aureus* where intensive vaccine research has seen many failures (Jansen et al., 2013).

The functional differences between the pan-genome components can be interpreted as follows: i) core genes are associated with functions related to the basic biology of the cell, such as central biochemistry and control and maintenance of gene expression machinery (housekeeping functions); ii) accessory genes are functionally similar to core genes, except for a relatively high proportion of genes associated with mobile genetic elements. Some of these genes are associated with lineage-specific replicons (i.e. plasmidic *rep* genes), which in turn may harbor resistance and virulence determinants (McCarthy and Lindsay, 2012); iii) unique genes are strongly associated with mobile elements. The lack of homology of unique genes with other *S. aureus* strains indicates that they have probably been acquired through horizontal gene transfer (HGT) events from relatively distant strains, or they are rapidly evolving genes such as those encoding outer membrane proteins.

It should be noted that our sampling of S. aureus strains for genome sequencing may be biased toward infections strains. Therefore it's possible that the dataset we selected is lacking commensal strains of *S. aureus* and more sequences of these strains would provide deeper insight into pathogenicity.

We found that *S. aureus* genomes are mostly composed of conserved genes (56% of genes in the average *S. aureus* genome are part of the core genome), with a very small proportion of strain-specific genes, thus revealing a high level of clonality for this species. It has been argued (Medini et al., 2005) that an open pan-genome is

typical of a taxon that can exchange genetic material with a variety of different sources. This, together with the fact that a greater portion of the unique genes have been assigned to COG category L (DNA recombination/repair), suggested that HGT is one of the major evolutionary forces shaping the genomic diversity of the *S. aureus* species. The impact of HGT was further investigated and we identified recent HGT-derived genes, their functional characterization and assigned a putative donor.

We assessed the impact of HGT events in shaping the diversity within this species. Analysis of atypical genes showed how these are mostly derived from taxonomically related donors (i.e. representatives of the same species/genus), with a minor portion of genes coming from host-associated bacteria. This finding suggests the presence of a taxonomical barrier limiting the amount of HGT in this species. The HGT events may be a major source of newly acquired antibiotic resistances. Therefore, the number of unique genes per genome may identify strains more prone to HGT and hence more likely to acquire new functionalities. Since virulence and antibiotic resistance genes are often involved in HGT events (McCarthy and Lindsay, 2012, Noble et al., 1992), strains with a higher portion of unique genes (and thus more predisposed to exchange of exogenous DNA) may represent a major public health threat, since these can develop virulent and/or multidrug-resistant phenotypes by horizontally acquiring corresponding gene cassettes (Corvaglia et al., 2010, Sung and Lindsay, 2007, Weigel et al., 2003) however we did not see any correlation between number of atypical genes and the number of identified virulence factors.

To gain insights into the metabolic diversity between *S. aureus* strains, we produced 64 strain-specific GEMs of our selected *S. aureus* strains and observed the presence of an extremely conserved metabolic core, both in terms of reactions and

genes. GEMs were analyzed to simulate *in silico* growth, to test alternative nutrient source usage and to compute genes essential for growth of *S. aureus*. Knowledge of essential genes may be used to guide future experiments aimed at finding molecular targets for the inhibition of pathogenic *S.aureus*. Moreover, we found a set of 148 essential genes shared between all the strains, representing potential broad-spectrum drug targets for this species.

Overall, most of the *S. aureus* virulome is conserved across the 64 strains we examined. There is only one VF unique to a single strain (i.e. the *etb* gene, encoding the exfoliative toxin B), while 54 VFs were shared by a small number of strains, but not present in all strains. The vast majority of the virulome is composed of core and pseudo-core (shared by most of the strains) genes. The presence of these genes is interesting from an evolutionary point of view, since it implies that the *S. aureus* has evolved a highly conserved system to carry out its infectious cycle, and from a translational viewpoint, these genes may represent targets for virulence inhibitor or antibody based therapeutic strategies (Morrison, 2015).

In conclusion, the multi-scale comparative approach used in this work allowed for deeper insights into the diversity of the *S. aureus* species. Results obtained from the pan-genome and the comparison of GEMs highlighted a low diversity at the genome-scale level, which is reflected both in small differences between GEM composition (in terms of genes, reactions, metabolites) and predicted phenotypes.

3.5 Acknowledgements

Chapter 3, in part, is a reprint of the material Bosi E*, Monk JM*, Aziz RK, Fondi M, Nizet V, Palsson BO. Comparative genome-scale modelling of multiple *S. aureus* strains identifies strain-specific pathogenic characteristics and unique metabolic

capabilities. *In Preparation.* The disseration author was the primary author (equally contributing with Emanuele Bosi) of this paper.

Figure 3.1. S. aureus dataset construction.
A) Phylogenic tree of 225 *S. aureus* genomes based on 7 housekeeping genes (*arcC, aroE, glpF, gmk, pta, tpi* and *yqiL)*. A set of 64 strains (labeled) were selected from this set to create a heterogeneous dataset of *S. aureus* strains based on **B)** drug resistance (MRSA, MSSA, VRSA, VISA), **C)** host specificity (human vs animal), **D)** virulence/environmental association (CA-MRSA: Community-Associated MRSA, HA-MRSA: Healthcare-Acquired MRSA, LA-MRSA: Livestock-Associated MRSA), iv) evolutionary distance based on tree topology.

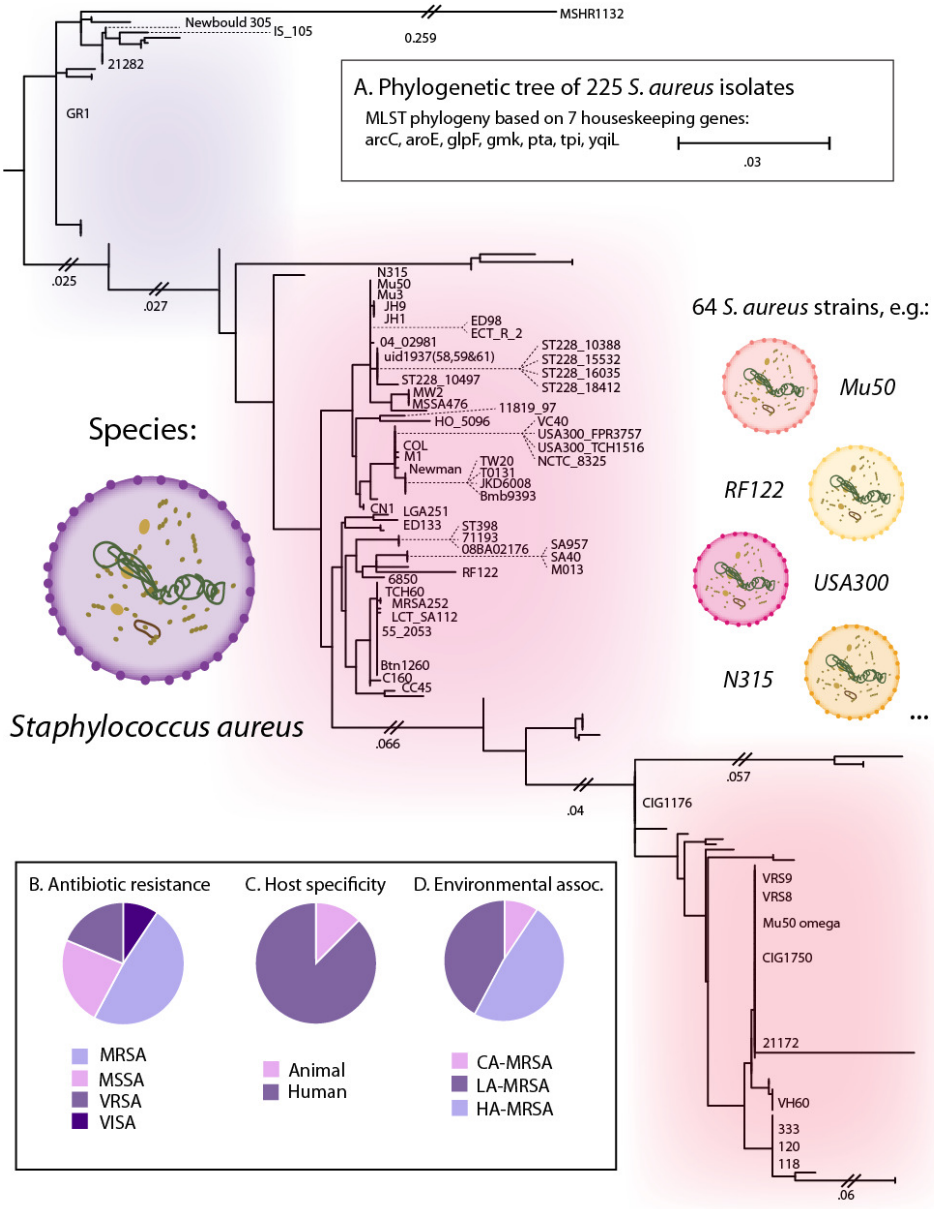
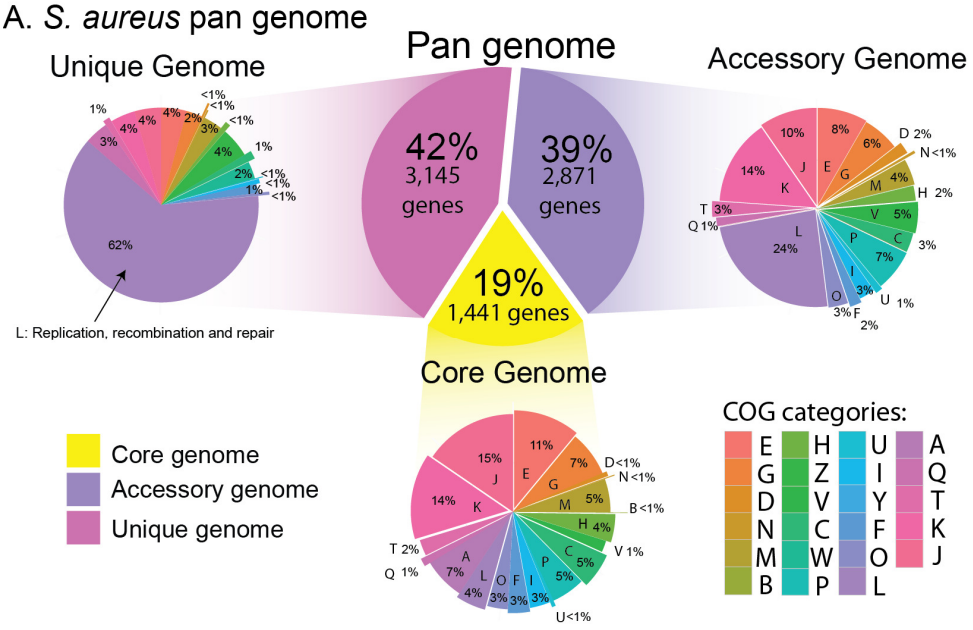
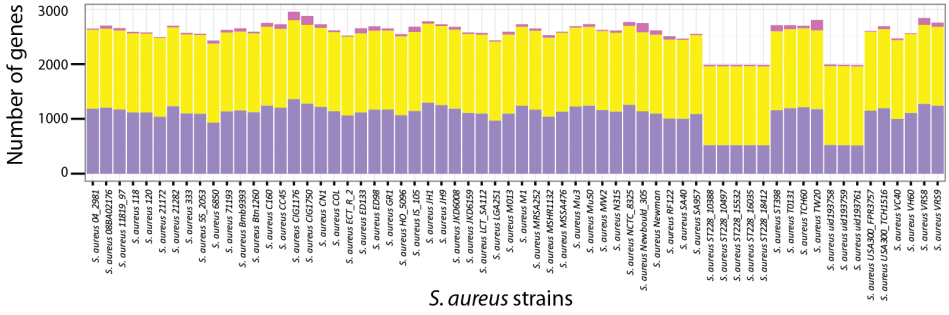

A. Phylogenetic tree of 225 S. aureus isolates
MLST phylogeny based on 7 houseskeeping genes:
arcC, aroE, glpF, gmk, pta, tpi, yqiL
.03
Newbould 305
IS_105
MSHR1132
0.259
21282
GR1
.025
.027
N315
Mu50
Mu3
JH9
JH1
ED98
ECT_R_2
04_02981
uid1937(58,59&61)
ST228_10388
ST228_15532
ST228_16035
ST228_18412
ST228_10497
MW2
MSSA476
11819_97
HO_5096
VC40
USA300_FPR3757
USA300_TCH1516
NCTC_8325
COL
M1
Newman
TW20
T0131
JKD6008
Bmb9393
CN1
LGA251
ED133
ST398
71193
08BA02176
SA957
SA40
M013
RF122
6850
TCH60
MRSA252
LCT_SA112
55_2053
Btn1260
C160
CC45
.066
.04
.057
CIG1176
VRS9
VRS8
Mu50 omega
CIG1750
21172
VH60
333
120
118
.06
Species:
Staphylococcus aureus
64 S. aureus strains, e.g.:
Mu50
RF122
USA300
N315
...
B. Antibiotic resistance
MRSA
MSSA
VRSA
VISA
C. Host specificity
Animal
Human
D. Environmental assoc.
CA-MRSA
LA-MRSA
HA-MRSA

Figure 3.2. *S. aureus* pan-genome statistics.
A) The S. aureus pan-genome can be subdivided into three categories: i) the core genome (the set of genes shared by all genomes), ii) the accessory genome (the set of genes present in some, but not all genomes) and iii) the unique genome (genes that are unique to a single genome). The function of each gene in a group is classified using clusters of orthologous croups (COGS). COG categories are as follows: Cellular processes and signaling: [D] Cell cycle control, cell division, chromosome partitioning, [M] Cell wall/membrane/envelope biogenesis, [N] Cell motility, [O] Post-translational modification, protein turnover, and chaperones, [T] Signal transduction mechanisms, [U] Intracellular trafficking, secretion, and vesicular transport, [V] Defense mechanisms, [W] Extracellular structures, [Y] Nuclear structure, [Z] Cytoskeleton. Information storage and processing: [A] RNA processing and modification, [B] Chromatin structure and dynamics, [J] Translation, ribosomal structure and biogenesis, [K] Transcription, [L] Replication, recombination and repair. Metabolism: [C] Energy production and conversion, [E] Amino acid transport and metabolism, [F] Nucleotide transport and metabolism, [G] Carbohydrate transport and metabolism, [H] Coenzyme transport and metabolism, [I] Lipid transport and metabolism, [P] Inorganic ion transport and metabolism, [Q] Secondary metabolites biosynthesis, transport, and catabolism. **B)** Distribution of the genes in the pan genome for each examined S. aureus strain.

A. *S. aureus* pan genome



B. Distribution of pangenome content in each *S. aureus* strain



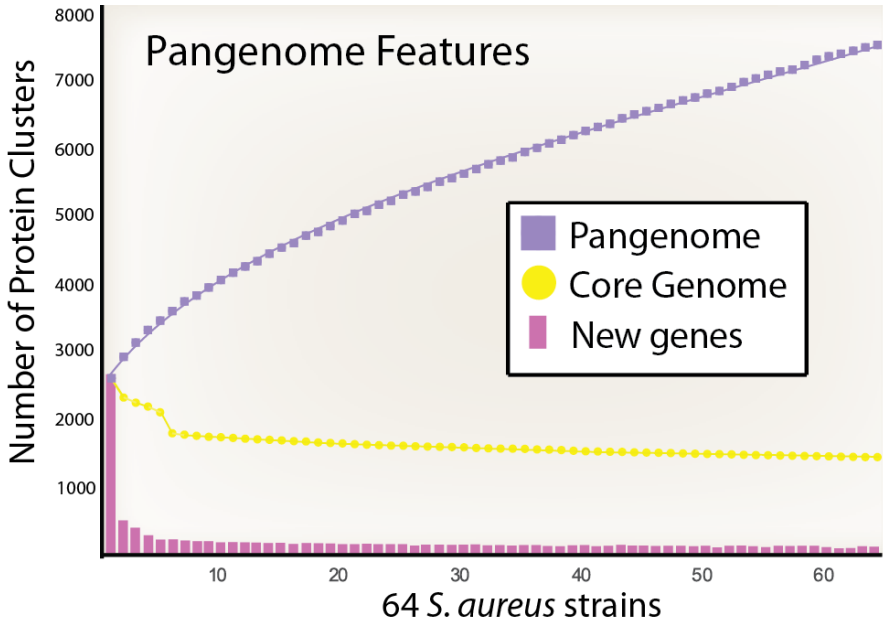


Figure 3.3. Pan-genome, core and novel genes of the 64 analyzed *S. aureus* strains. Pangenome features: the purple squares denote the number of novel genes discovered with the sequential addition of new genomes. The yellow dots denote the values of the core genes as genomes are added to the pangenome. The purple bars indicate the number of new genes added to the total pan genome size as new genomes are added. Each of the values represents the median from a distribution of randomly selected genomes at each genome addition. The purple line represents the number of new genes found for each genome addition. For comparison, the same trend for a closed genome is reported as a dashed line.

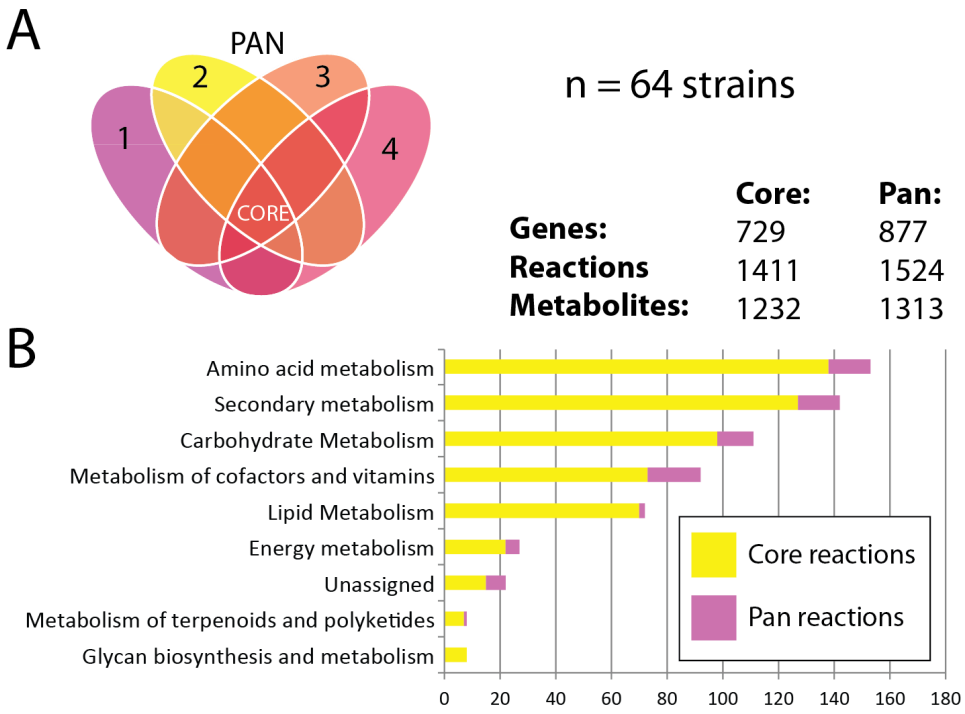


Figure 3.4. Core and Pan metabolic capabilities of the *S. aureus* species.
The core and pan metabolic content was determined for genome-scale metabolic models (GEMs) of 64 unique *S. aureus* strains. **A)** The core content, illustrated by the intersection of the Venn diagram, is shared with all strains. The pan content consists of all content in any model and includes the core content. The Venn diagram is not to scale. **B)** Classification of reactions in the core and pan reactomes by metabolic subsystem.

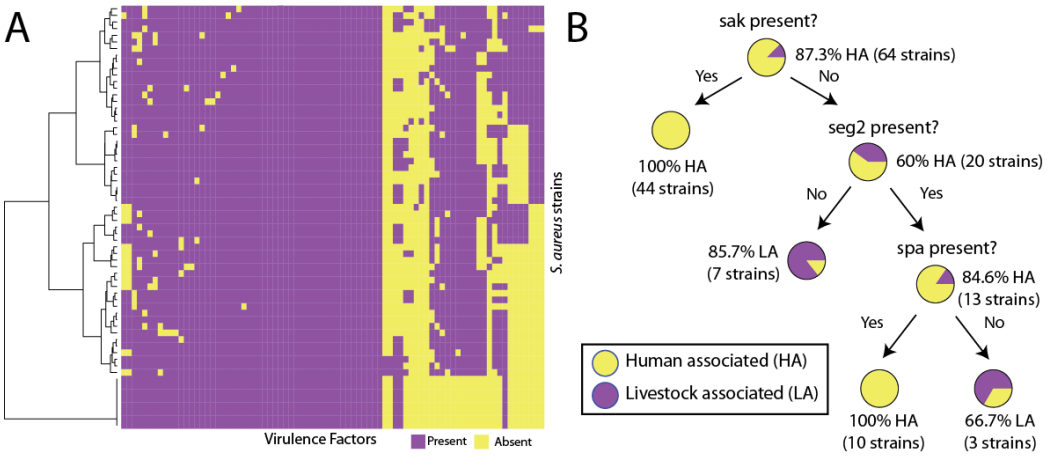


Figure 3.5. *S. aureus* virulome.
The virulome consists of curated virulence factors known to be present in different strains of *S. aureus*. **A)** Presence and absence of these *S. aureus* virulence factors across the strains examined in this study. Purple = present, Yellow=absent. Full matrix with strains and virulence factor is available in the supplement. Virulence factor profiles can be used to classify strains, for example in (**B)** a classification is constructed that separates human associated *S. aureus* strains from livestock associated strains using the presence of three specific virulence factors. Abbreviations: Staphylokinase precursor (sak), staphylococcal enterotoxin (seg2), Immunoglobulin G binding protein A precursor (spa).

Chapter 4 Comparative metabolic network analysis and modelling of four Leptospira species provides insight into pathogenesis of Leptospirosis

4.1 Abstract

Multiple Leptospira genome sequences and new omics data have recently been made available by advances in DNA sequencing. Analysis of these genomes shows that the fraction of genes common to all Leptospira represents a small fraction of the entire Leptospire gene pool. Thus the question arises: what makes a species of Leptospira pathogenic? We constructed genome-scale models of four Leptospira species that ranged in their level of pathogenicity. From the pathogenic *L. interrogans*, to intermediate species *L. kmetyi* and *L. licerisae* to the non-pathogenic, saprophyte, *L. biflexa*. The GEMs enable a systems approach to characterizing the core and pan metabolic capabilities of the Leptospira genus. The majority of shared metabolic content was found to consist of lipid metabolism, energy production and conversion and amino acid metabolism, while reactions unique to specific species were often found in carbohydrate metabolism, nucleotide metabolism and cofactor and prosthetic group metabolism. The results show that unique species-specific metabolic capabilities correspond to pathotypes and environmental niches. All four species were predicted to be auxotrophic for L-asparagine and vitamin B1. Cob(1)Alamin (vitamin b12) was required for growth of the saprophytic strain but not for the other strains. GEMs were used to predict shared and species-specific essential genes and reactions. The Leptospira genus is predicted to have 264 universally essential metabolic reactions and 112 reactions that are essential to sets of Leptospira species. A total of 366 metabolites were predicted to be essential across the Leptospira genus with 44 that were specific to specifically essential to sets of Leptospira species. These results indicate that species specific drug targets and metabolite analog inhibitors may be capable of selectively targeting individual species of or groups of Leptospira (e.g. pathogens). Genome-scale

analysis of multiple species in a genus can thus be used to define the metabolic essence of a microbial genus and delineate nutrient and essentiality differences that shed light on the metabolic determinants of pathogenicity.

4.2 Introduction

Leptospira is a genus of highly motile gram negative spirochetes capable of penetrating the mucous membrane, eyes and abraded skin to cause systemic disease (Bharti et al., 2003). Several species of *Leptospira* are the causative agents of leptospirosis, a zoonotic disease that affects nearly one million people per year in countries worldwide (Lau et al., 2010). Live imaging of bioluminescent cells of *Leptospira interrogans*, the best characterized leptospiral species, revealed that, once inside the host, these pathogenic leptospires can evade the host immune system and antibiotic treatment by colonizing the kidneys (Ratet et al., 2014). Patients diagnosed with severe leptospirosis, also known as Weil's disease, can suffer from jaundice, renal failure and pulmonary hemorrhage. This disease is endemic in countries with tropical climates, often as a result of floods caused by heavy rain (Saito et al., 2014, Lau et al., 2010). Because the disease is transmitted through the infected urine of mammalian carriers either directly or via contamination of water or soil, individuals living with poor sanitary conditions are more frequently infected.

Worldwide, leptospirosis is one of the most common diseases transmitted by animals. It is a major cause of illness in tropical areas and is often associated with epidemics during natural disasters and flooding. For this reason, climate change is expected to increase the incidence of leptospirosis infection in tropical regions worldwide. Leptospirosis is often unrecognized or misdiagnosed due to its symptoms that are similar to many other diseases. These symptoms frequently present themselves

as fever or myalgia, making the actual incidence of leptospirosis underreported. Further complicating matters, laboratory tests used to detect and confirm the disease are not always available in developing countries (WHO, 2003). For these reasons, Leptospirosis has been recognized as a neglected tropical disease of global importance that necessitates further research.

The *Leptospira* genus has both pathogenic and saprophytic members. More than twenty species of *Leptospira* have been described and grouped into three distinct lineages. These are 1) pathogenic (nine species), 2) intermediate (five species) and 3) saprophytic (non-infectious) (six species) (Ricaldi et al., 2012). Pathogenic and intermediate species such as *L. interrogans*, *L. Kmetyi* and *L. licerasiae* can infect a wide range of animals, including rats, domestic pets, and humans (Lau et al., 2010). Saprophytic strains such as *L. biflexa* cannot infect humans (Picardeau et al., 2008).

The amount of literature regarding Leptospiral organisms has increased in recent years reflecting the growing rate of leptospirosis worldwide (**Figure 4.1A**) and a newfound desire to investigate and combat this epidemic. Furthermore, advances in DNA sequencing technologies have enabled the generation of new -omics data types (genomic, proteomic, and transcriptomic) for *Leptospira* in the last decade. However, as useful as this new data and genome annotations are, they do not provide an understanding of the integrated function of gene products to produce phenotypic states (Bordbar et al., 2014). Metabolic network reconstructions have proven to be powerful tools to probe the genomic diversity of metabolism between organisms (Monk et al., 2013). Despite an increase in Leptospira-specific data types, no genome-scale network reconstructions (GENREs) of any *Leptospira* species exists. Indeed, the entire

spirochetes phylum is void of any representative GENRE at this time (Monk et al., 2014b) (**Figure 4.1B**). Thus, constructing new metabolic network reconstructions of *Leptospira* species will increase biochemical insight into a unique branch of the tree of life.

We set out to construct metabolic network reconstructions of a set of representative species within the *Leptospira* genus by selecting species that span a range of infectious capabilities from the extremely infectious (*L. interrogans* serovar Copenhageni Fiocruz L1-130) to the intermediate-infectious strains (*L. kmetyi* serovar Malaysia Bejo-Iso9 and *L. licerasiae VAR010*) to the non-infectious, saprophytic (*L. biflexa*).

L. interrogans is responsible for worldwide cases of the waterborne zoonosis leptospirosis. It infects wild and domestic animals, including pet dogs. It is usually transmitted to humans via contact with infected animal urine where it invades directly through broken skin and can infect the kidney and liver. The two intermediate leptospires, *L. kmetyi* and *L. licerasiae* display some phenotypic characteristics of saprophytic leptospires, however 16S rRNA gene sequence analysis suggests that they are more closely positioned phylogenetically among pathogenic Leptospires and a whole genome analysis supports this conclusion as well (Ricaldi et al., 2012). *L. licerasiae* has also been proposed to have proteins involved in nitrogen, amino acid and carbohydrate metabolism unique to the genus, which could explain *L. licerasiae*'s quick growth in artificial media, reportedly posing contamination problems for biopharmaceutical groups (Chen et al., 2012a).

The non-pathogenic, saprophytic leptospire, *L. biflexa,* is found in aquatic environment. In contrast to many pathogenic leptospires, the genome of free-living *L.*

biflexa contains high gene density and few transposable elements, key features for the study of laterally transferred genes in *Leptospira* (Picardeau et al., 2008). Comparative genome analyses have shown that nearly one third of *L. biflexa* genes are absent from pathogenic leptospires. Thus, studies elucidating the gene functions present in *L. biflexa*, but absent in pathogenic leptospires can provide insight regarding the evolution of leptospiral species.

In this study, we present curated metabolic network reconstructions of these four species of *Leptospira* displaying a range of infectious phenotypes. The reconstructions incorporate available literature and experimental data on the *Leptospira* species and are used to investigate the metabolic capabilities that may be prerequisites of pathogenicity.

4.3. Results

4.3.1 Core Leptospira metabolism

The four metabolic network reconstructions were built following an established protocol (Thiele and Palsson, 2010b). Annotations of the genomes were downloaded from NCBI and functional assignment and metabolic content were curated from metabolic databases including model seed (Devoid et al., 2013, Henry et al., 2010c) Metacyc (Caspi et al., 2014) and Metanetx (Ganter et al., 2013). An analysis of transport proteins related to metabolic intermediates was performed based on a recent study (Buyuktimkin and Saier, 2015). The detailed list of the contents of each model is available in **Supplementary Dataset 1**.

The set of four *Leptospira* genome-scale reconstructions were used to compare gene, reaction, and metabolite content between species (**Figure 4.2A**). The content shared among all reconstructions thereby defines the core metabolic capabilities among all of the species. Similarly, the metabolic capabilities of all of the species were

combined to define the full set that encompasses all models and thereby define the “pan” metabolic capabilities among the species.

The size and content of the core metabolic content characterizes the metabolic foundation of the *Leptospira* genus. The core reconstruction has 581 metabolic genes that catalyze 943 reactions using 736 metabolites. The most highly conserved metabolic subsystems were lipid metabolism (88% core), amino acid metabolism (82% core) and cell wall, membrane and envelope metabolism (78% core). Most of these reactions are responsible for synthesizing essential components fatty acids, cell wall components and amino acids. In contrast, only 63% of carbohydrate metabolism reactions were shared among all four species. These shared reactions mostly consisted of reactions in central metabolism reactions including glycolysis/gluconeogenesis, the citric acid cycle, anaplerotic reactions, and the pentose phosphate pathway.

The core metabolic content contains the complete set of genes for the Tricarboxylic Acid Cycle (TCA) and respiratory electron transport chain (Ren et al., 2003, Nascimento et al., 2004a). In *L. interrogans*, genes associated with the TCA cycle have been shown to be expressed at similar levels under *in vitro* conditions (EMJH medium at 30C) and mammalian host conditions (dialysis membrane chamber, or DMC, within the peritoneal cavities of *R. norvegicus*) (Caimano et al., 2014). ATP generated by the respiratory electron transport chain uses an F0F1-type ATPase that is encoded in a single operon, atpBEFHAGDC, and shared among leptospires (BIGG: ATPasel) (Ren et al., 2003).

All genes required for glycolysis and glucose uptake are also represented in the core metabolic content. RNA-seq experiments have shown that reads for all genes involved in glucose metabolism and transport were detected and expressed at similar

levels in mammalian host conditions (DMC) and *in vitro* conditions (EMJH medium) (Caimano et al., 2014). A glucokinase gene is proposed to substitute for a hexokinase gene in leptospires (Picardeau et al., 2008, Zhang et al., 2011). While many Leptospira species do contain these genes, pathogenic leptospires are proposed to not utilize glucose from the environment (Baseman and Cox, 1969). Only the enzymes that catalyze reversible reactions (all those except glucose-6-phosphate isomerase (PGI) and pyruvate kinase (PYK)) have been shown to be expressed at high levels in *L. interrogans*, which provides support for gluconeogenesis activity, but not glycolytic activity. Additionally, the two genes, LIC13358 (bigg: RZ5PP) and LIC20119 (bigg: PGM), encoding putative phosphoglucomutases, and LIC12908 (sglT; FRUt4pp; GALt4pp), encoding the only glucose transporter identified in *L. interrogans* (Nascimento et al., 2004b), were expressed at extremely low levels under both mammalian host conditions and *in vitro* conditions (Caimano et al., 2014). These experimental findings suggest that utilization of glucose is unlikely.

Finally, the genes encoding enzymes in the non-oxidative phase of the pentose phosphate pathway are all present in the core genome. Only one gene (LIC12161) for the oxidative phase encoding 6-phosphogluconolactonase (PGL) is represented in the core genome, supporting the claim that leptospires use an alternative pathway, (possibly the the NAD+ salvage pathway) to generate NADPH (Qu, 2007).

4.3.2 Pathways unique to core *Leptospira* metabolism

Beyond these canonical metabolic pathways that are widely conserved across bacteria, the core Leptospira metabolic reconstruction also contained several interesting metabolic pathways that are less frequently observed in gram negative bacteria. All four of the Leptospira organisms possess an alternative menaquinone biosynthesis

pathway. In most bacteria menaquinone is an essential vitamin that is a required component of the electron transport chain. In *E. coli* menaquinone is derived from chorismate by eight enzymes, MenA-H (Bentley and Meganathan, 1982). Bioinformatic analysis of the four leptospira species showed a lack of orthologs to the *men* genes. Instead orthologs were found that correspond to an alternate menaquinone pathway originally characterized in *Streptomyces coelicolor* known as the futalosine pathway (Seto et al., 2008, Hiratsuka et al., 2008) (**Figure 4.3A**). Furthermore, all 4 species were found to possess transporters for 4-Hydroxybenzoate, which acts as an intermediate in this biosynthesis pathway. This pathway is also found in several pathogenic organisms including Helicobacter pylori and Campylobacter jejuni (Dairi, 2009), however is missing in humans and other commensal bacteria, making the enzymes in this pathway promising antibiotic targets that could specifically target growth of pathogens like Leptospira with minimal effect on the intestinal microflora.

Another essential vitamin for cellular growth is folate (vitamin B9). Analysis of the core Leptospira metabolic content indicates that all four strains of Leptospira can synthesize folate *de novo*, however they possess a unique dihydroneopterin aldolase (FolB) enzyme. FolB catalyzes the conversion of dihydroneopterin to 6-hydroxymethyl dihydropterin (HMDMP) in the classical folate biosynthesis pathway, however standard folB genes are missing in several phyla including the *Spirochaetes* (de Crecy-Lagard et al., 2007). These organisms, including the four *Leptospira* examined here instead possess an unusual paralog of the tetrahydrobiopterin synthesis enzyme 6-pyruvoyltetrahydropterin synthase (PTPS) (Pribat et al., 2009) also making this enzyme a potential target for novel antifolate compounds.

Lysine biosynthesis in Leptospira occurs via a pathway that utilizes diaminopimelic acid (DAP) as a central intermediate. However Leptospires utilize an alternative lysine biosynthesis pathway shared with cyanobacteria that utilizes a novel transaminase to catalyze the interconversion of tetrahydrodipicolinate and LL-diaminopimelate. Thus accomplishing in a single step a conversion that requires three enzymes in the DAP-pathway variant found in *Escherichia coli* (Hudson et al., 2006). Also, methionine biosynthesis in leptospires is reportedly similar to that in yeast, where the final step has been proposed to be exclusively catalyzed by a cobalamin-dependent homocysteine- N5-methyltetrahydrofolate transmethylase (MetH), rather than by a cobalamin-independent methionine synthase (MetE) (Ren et al., 2003). The four models in this study are shown to contain the *metH* gene (KEGG R00946; LIC20085; LBF_4176; LEP1GSC185_3695; LEP1GSC052_0294 → bigg: METSr), and lack the *metE* gene (bigg: MHPGLUT, MTHPTGHMr) thus making them dependent on cobalamin for methionine synthesis.

Genomes of leptospires contain the complete set of genes necessary for protoheme uptake and biosynthesis (Guegan et al., 2003, Murray et al., 2009). Sequence analysis of chromosome II shows that an almost complete cluster of genes coding for the protoheme biosynthesis pathway is present (*hemAIBCENYH*), and another gene, *hbpA*, has been characterized to encode a hemin-binding protein that is expressed under iron starvation. Although no homologue of the gene coding for uroporphyrinogen III synthetase (*hemD*) has been found, experimental evidence has shown that the *hemC* gene contains *hemD* activity (Guegan et al., 2003). Therefore all of the Leptospires examined here are predicted to have the ability to synthesize protoheme *de novo*. RNA-seq studies with *L. interrogans* have demonstrated that 5 genes related

to *de novo* heme biosynthesis (LIC20008/*hemA*, LIC20009/*hemCD*, LIC20010/*hemB*, LIC20011/*hemL* and LIC20014/*hemE*) were downregulated under mammalian host conditions (DMCs) compared to *in vitro* (EMJH) conditions. In *L. interrogans, the* LIC20017/*hemG*/*hemeY* and LIC20018/*hemH*, encoding enzymes responsible for the last two steps in heme biosynthesis, respectively, appear to be transcribed as monocistronic messages at similar levels *in vitro* (EMJH) and under mammalian host conditions (DMCs) (Caimano et al., 2014). Additionally, genes for heme oxygenase (LIC20148/*hol*) and a TonB-dependent heme receptor (LIC10964/*phuR*) were upregulated under mammalian host conditions (DMCs) compared to *in vitro* (EMJH). These data support the notion that pathogenic leptospires preferentially use exogenously derived heme within the mammal (Caimano et al., 2014). And further suggest that protoheme biosynthesis in pathogenic leptospires is reduced when scavenging protoheme from the mammalian host is possible (Caimano et al., 2014).

Such scavenging occurs via the high-affinity TonB-dependent outer membrane receptor (TB-DR) proteins. Based on bioinformatic analysis, *L. interrogans* encodes at least 13 putative TB-DRs (Louvel et al., 2006), however, only two (LIC10964 and LIC11694) were upregulated within DMCs (Caimano et al., 2014). The transport of heme and/or iron across the outer membrane requires energy produced by an inner membrane complex of the energy transduction protein TonB and two accessory proteins, ExbB and ExbD (Noinaj et al., 2010). *L. interrogans* encodes at least two TonB-ExbB-ExbD complexes, arranged in separate operons, one on each chromosome (Caimano et al., 2014). *Leptospira* species have been shown to possess multiple outer membrane energizer systems (e.g TonB-ExbB-ExbD and TolA-TolQ-TolR) (Buyuktimkin and Saier,

2015) with the latter facilitating assembly and stability of the outer membrane (Lazzaroni et al., 1999).

4.3.3 Pan *Letospira* metabolism

The pan metabolic reconstruction of the Leptospira genus encodes complete pathways for most amino acids and all nucleic acid biosynthesis. However, genomes of *L. licerasiae* and *L. biflexa* encode additional proteins involved in nitrogen, amino acid, and carbohydrate metabolism that are missing from pathogenic Leptospires. This may account for the ease of growth of these species in vitro (Ricaldi et al., 2012). *L. licerasiae* (LEP1GSC185_2652) and *L. biflexa* (LEPBI_I1590; LBF_1539) both possess *ilvA* (bigg:THRD_L; also bigg:SERD_L), which encodes threonine ammonia-lyase, an enzyme that catalyzes the conversion of threonine to 2-oxobutanoate. *L. interrogans* and *L. kmetyi* do not contain the *ilvA* gene.

In addition to unique amino acid metabolism, only 74% of cofactor and prosthetic group metabolism reactions were shared among all four species. Reactions in this group are responsible for biosynthesis of essential vitamins including vitamins B12 (cobalamin), B9 (folate) and B1 (thiamin). A major difference was found to be presence/absence of the cobalamin (vitamin B12) biosynthetic pathway. Cobalamin is the largest and most complex of natural organometallic cofactors and coenzymes, its *de novo* synthesis requires nearly 30 energetically costly enzymatic steps (Raux et al., 2000). Two distinct biosynthetic pathways exist, which are referred to as the aerobic (Heldt et al., 2005) and anaerobic (Moore and Warren, 2012) routes (**Figure 4.3A**) that differ based upon the timing of cobalt ion insertion into the corrin ring and specific genes needed. Our analyses indicate that pathogenic *Leptospira* contain genes necessary to synthesize B12 *de novo* while the saphrophyric *L. Biflexa*, does not. All of the metabolic

reactions, genes and metabolites discussed in this manuscript can be visualized and compared between all four species using the BiGG database at http://bigg.ucsd.edu (King et al., 2015c) (**Figure 4.3B,C**).

Previous studies have identified 13 genes clustered in chromosome II coding for the cobalamin biosynthesis pathway *(cobC, cobD, cbiP, cobP, cobB, cobO, cobM, cobJ, cbiG, cobI, cobL, cobH, cobF*) (Nascimento et al., 2004b). Orthologues of *cobGKN* genes, known to function as cobalt chelatases in the cobalamin pathway (Rodionov et al., 2003), were not found. However, two predicted coding sequences inside this operon in chromosome II have been proposed to perform these steps. One has an oxidoreductase NAD-binding domain (LIC20133) and the other is a [2Fe-2S] ferredoxin involved in electron transfer (LIC20135) (Nascimento et al., 2004b). In addition, other genes present in the genome coding for reductases such as LIC11145, LIC13354, LIC12391, and LIC10522 could also fulfill these activities (Spencer et al., 1993). The presence of *cysG* in chromosome I, a gene that encodes a multifunctional protein with methylase, oxidase and ferrochelatase activities, may function as a cobalt-inserting enzyme in the B12 pathway. Other genes involved in this biosynthesis pathway were found in chromosome I *(cysG/hemX/cobA, cobT/cobU, cobS*). Further, the presence of *cbiX* (LB_165) and *cbiG* [DEF1] (LB_158) in cob I/III suggests that infectious *Leptospira* utilize the anaerobic biosynthetic route for cobalamin biosynthesis. It was recently experimentally confirmed that *L. interrogans Copenhageni* PL1, *L. licerasiae Varilla* ***I*** *VAR010* grow indefinitely in media without B12 while the saprophyte *L. biflexa* Patoc Patoc1 (Paris) does not (manuscript in preparation – Vinetz group). Our analysis therefore predicts that *L. biflexa* requires B12 for growth. All four species are predicted

to possess vitamin B12 transporters and a TB-DR system in *L. interrogans sv. Copenhageni* (LIC12374/btuB) is annotated as being specific for vitamin B12.

Though only infectious *Leptospira* contained the genes necessary to synthesize B12 *de novo*, both infectious and non-infectious *Leptospira* contain reactions necessary to synthesize the vitamin from the intermediate adenosylcobinamide (through an ortholog for CobU, which converts adenosylcobinamide to andeosylcobinamide-guanosine diphosphate, has not been found in non-infectious species). Why only infectious and not non-infectious *Leptospira* synthesize B12 *de novo* is not known, but considering that cob I/III clusters are widely distributed amongst pathogenic strains and that in mammals the vitamin is sequestered *in vivo*, it is likely that this capacity is essential for growth *in vivo*. Therefore B12 biosynthesis may be an important virulence trait.

4.3.4 Metabolic models reveal species-specific nutrient requirements

Based on the analysis of metabolic content described above, we set out to investigate the required nutrients for growth of each species of *Leptospira*. Converting a genome scale reconstruction into a mathematical format, called a Genome-scale Model (GEM) allows for explicit computation of growth capabilities and the examination of nutrients required for growth in different environments that can be linked back to their genetic basis. The first step in converting a genome scale network reconstruction into a genome-scale model is to define a biomass objective function (Feist and Palsson, 2010b). The biomass objective function consists of all nucleic acids, amino acids, lipids etc. needed for a cell to grow (and produce biomass). The biomass objective function must be determined from measurements of biomass composition. A detailed analysis of Leptospira biomass has not been performed, however coarse grain measurements

exist. We tailored a generic gram-negative biomass function to match these measurements to form a core biomass function representative of the Leptospira genus (**Table 4.1**).

None of the four reconstructed GEMs were able to generate essential biomass components from glucose M9 minimal media without addition of growth-supporting compounds to the in-silico media. The SMILEY algorithm, a method to fill gaps in metabolic networks (Reed et al., 2006a), was used to examine the genetic bases of these model nutrient requirements. We defined a minimal media for growth of Leptospira that supports the growth of each species of leptospira (**Table 4.2**).

All models predicted that fatty acids were required for growth. It is known that Leptospira are unable to synthesize fatty acids from pyruvate or acetate due to an absence of genes involved in fatty acid biosynthesis chain elongation (Johnson et al., 1970, Stern et al., 1969). However, the complete set of genes for beta-oxidation are present. Therefore, fatty acids must be obtained through a given growth medium or from the solid-liquid interfaces in natural environments where fatty acids are located (Kefford and Marshall, 1984). *L. biflexa* growth can be supported on long or short, saturated or unsaturated FAs (Johnson et al., 1969, Khisamov and Morozova, 1988) while *L. interrogans* requires the addition of long chain unsaturated FAs to be added to the growth medium in order to synthesize saturated FAs. Also, *in vitro* growth of leptospires is greatly improved with the addition of glycerol as an additional carbon source beyond the fatty acids mentioned here (Staneck et al., 1973). This observation is recapitulated by our models where addition of glycerol to the *in-silico* media increases predicted growth rates.

All models were capable of synthesizing all 20 amino acids except for asparagine. This was determined to be due to a lack of asparagine synthase (EC: 6.3.1.1, ASNS1). It has previously been demonstrated that growth of Leptospira pomona is stimulated upon addition of L-asparagine to the media and L-asparagine was the only amino acid that markedly stimulated growth (Johnson and Gary, 1962). Furthermore, L-asparagine has long been recognized as a standard addition to Leptospira Medium.

Metabolic models predicted also predicted auxotrophies for essential vitamins. All species were predicted to require thiamin (vitamin B1) for growth due to a lack of thiazole synthase (THZPSN, EC: 2.8.1.10). Thiamin auxotrophy is a well-documented feature of leptospires. (Faine, 1959, Staneck et al., 1973). Furthermore, vitamin B12 was predicted to be required for growth of *L. Biflexa*. This is because it lacks the genes required for B12 biosynthesis (described above). All other models of Leptospira did not require B12 to be added to the *in-silico* minimal media indicating that they were capable of synthesizing this vitamin *de-novo.*

The metabolic models predicted that all strains require a nitrogen source which was provided in the form of ammonium (Faine, 1999). However, the pathogenic species were predicted to be capable of substituting ammonia with urea due to the presence of urease (EC: 3.5.1.5) which hydrolyzes urea into carbon dioxide and ammonium for use a nitrogen source. It has previously been observed that one representative of each of five different pathogenic serotypes of Leptospira were capable of growing on medium containing urea in place of an ammonium salt as a nitrogen source (Kadis and Pugh, 1974).

4.3.5 Using genome-scale models to predict essential reactions and metabolites

Genome scale models have been used to predict essential genes and reactions in several organisms (Monk and Palsson, 2014) with up to 90% accuracy for well-studied organisms (Orth et al., 2011). Using the minimal media formulation defined above, we used the models to systematically probe *Leptospira* metabolism using a systematic deletion of all metabolic reactions encoded in each model (**Figure 4.4A**). The model for *L. interrogans* had the most predicted essential reactions (335) followed by *L. licerasiae* (327), *L. Kmetyi (318)*, and *L. biflexa (313)*. Of these, the four models shared 264 essential metabolic reactions (**Figure 4.4B**), while 112 reactions were uniquely essential to specific sets of species (**Figure 4.4C**). The largest group of shared essential reactions fell into the metabolic subsystem of cofactor and prosthetic group biosynthesis (70 reactions), followed by amino acid metabolism (59 reactions) and cell envelope biosynthesis (20 reactions). The unique essential reactions fell into different subgroups of species. For example the three pathogenic leptospires shared 21 essential reactions that were dispensable in the saprophyte *L. biflexa* while 7 reactions were predicted to be specifically essential to *L. biflexa* but dispensable in the other three species. The full list of model-predicted universally essential reactions and their catalyzing genes in each species is available in **Supplementary Data File 2.**

In addition to an analysis of essential reactions, metabolic models can be used to study essential metabolites by removing their consuming reactions from the model, thereby predicting the effect of potentially growth inhibiting metabolite analogues (Kim et al., 2010b) (**Figure 4.4D**). We predicted essential metabolites for each species using the metabolic models and found that *L. Interrogans* had the most predicted essential metabolites (397) followed by *L. kmetyi* (386), *L. licerasiae* (384) and *L. biflexa* (383) for an average of 390 essential metabolites out of a total average of 950 metabolites in

each model (41%) this is in line with other organisms like *H. pylori* (40.9%) , but considerably more essential metabolites than for *E. coli* (24.8%) (Kim et al., 2007) and *V. vulnificus* (25.2%) (Kim et al., 2011). However those studies examined essential metabolites using an in-silico complex media that included all amino acids, unlike the minimal media defined here which may have caused a larger number of essential metabolites to be predicted in the case presented here. Among the essential metabolites we predicted here, 366 of them were shared among the four species and 44 were unique to different sets of species. For example, four metabolites were predicted to be specifically essential to *L. interrogans*.

The analysis of essential metabolites relies on removing reactions based on their consumption of a particular metabolite. Therefore it is similar to the essential reaction analysis conducted above but also accounts for multiple consuming reaction deletions that may result due to the presence of a metabolite inhibitor. Given the genetic intractability of several Leptospira species (Picardeau et al., 2001, Girons et al., 2000) using predicted essential metabolites to experimentally test Leptospira metabolism may be easier than exploring essential reactions (and knockout of their catalyzing genes). Indeed specific metabolite analog inhibition of Leptospires has already been documented as a way to distinguish species of Leptospira. Johnson and Rogers demonstrated in 1964 that the guanine analog 8-azaguanine specifically inhibited the growth of pathogenic Leptospires while Saprophytic species were almost unaffected (Johnson and Rogers, 1964). Therefore the *in-silico* predictions of metabolite inhibitors presented here might be a useful approach to delineate species or groups of Leptospira (eg. pathogenic species) for rapid identification and classification of individual Leptispira species based on their response to specific chemical inhibitors.

4.4 Discussion

The study of the Leptospira genus and its species-specific differences has progressed significantly since Edward Hindle found in 1925 that leptospires isolated from London tap water would grow in feces medium while pathogenic strains would not (Hindle, 1925). Modern advances in DNA sequencing technologies and microbiology techniques have made the study of this difficult to culture organism tractable. Here we report the first genomically-predicted metabolic network analysis (O'Brien et al., 2015) of *Leptospira,* comparing members of the pathogen, intermediate pathogen and saprophyte clades. These reconstructions allow comparison of the conserved metabolic capabilities (core metabolic network) and the unique metabolic capabilities (pan metabolic network) for the Leptospira genus. Leptospira core metabolism is unique and distinct from other gram-negative organisms such as *E. coli* in several ways including a unique menaquinone biosynthesis pathway, alternate enzymes involved in folate metabolism, a unique lysine biosynthesis pathway and more.

The most striking difference between the metabolism of the infectious *Leptospira* species examined and the non-pathogenic species *L. biflexa* was found in cofactor and vitamin biosynthetic capabilities. The models of *L. interrogans, L. L. Kmetyi and L. licerasiae* have a complete vitamin B12 biosynthetic pathway that enables *de novo* B12 synthesis from an L-glutamate precursor, while *L. biflexa* completely lacked this pathway. These differences in biosynthetic capabilities may allow such pathogens to survive in nutrient-limited niches within the human body. These observations are consistent with previous observations that found that the pathogenic strain *L. interrogans* serovar Canicola can grow *in vitro* in the absence of vitamin B12 but not vitamin B1 (Stalheim and Wilson, 1964). In addition to biotin, our analyses identified urea utilization

as an important factor differentiating the pathogenic *Leptospira interrogans* from the other species. Pathogenic leptospires can cause renal failure (Yang et al., 2001) and other studies have suggested that highly active urease influences the persistence of pathogens in the kidney (Braude and Siemienski, 1960, Lovell and Harvey, 1950).

We used the metabolic models generated here to predict a minimal media capable of supporting of growth for all of the Leptospires. The media required L-arginine, thiamin, cobalamin (for *L. biflexa*) and medium and long chain fatty acids along with nitrogen, phosphorous, sulfur sources and essential ions to support growth. This predicted media is consistent with established Leptospira culture conditions. Next we used the models to predict essential reactions and metabolites for each species of Leptospira. This analysis demonstrated that a large set of essential functions is shared between species of Leptospira is (70% of reactions and 88% metabolites), but that species-specific essentialities exist. Such knowledge represents potentially new species-specific drug targets as well as new ways to test, classify and identify species of Leptospira among the genus.

Metabolic models can also be used to answer longstanding questions of Leptospira biology. One such questions relates to why *L. interrogans* grows more slowly than do intermediate pathogens and saprophytes, such as *L. licerasiae* and *L. biflexa*, which grow rapidly in defined EMJH media (Matthias et al., 2008). The metabolic network model of *L. interrogans* was shown to lack L-glutamate oxidoreductase, an enzyme involved in recruiting ammonia as a nitrogen source (Murachi and Tabata, 1987, Bohmer et al., 1989), predicting a lower growth yield compared to the other *Leptospira* models in our *in-silico* minimal media analysis. The model of *L. biflexa* predicted the greatest yield with this reaction because this *Leptospira* contains L-aspartate ammonia-

lyase, allowing it to convert L-aspartate into fumarate and ammonia, in addition being capable of using this nutrient for biomass generation. These observations could hint at one possible solution to the question of different growth rates, but further model-guided experimentation is required to validate this prediction.

Further curation and experimental validation of the four metabolic models presented here for members of the Leptospira genus will open reconstruction opportunities for other related organisms in the spirochetes phyla including *Borrelia* spp., which cause Lyme borreliosis and relapsing fever, and *Treponema* spp., which cause syphilis, yaws, periodontitis and other diseases. Such an approach will yield new, fundamental insights into the diverse metabolic capabilities of this phylum.

4.5 Materials and Methods

4.5.1 Metabolic network reconstruction procedure

The genome-scale metabolic networks were reconstructed according to an established protocol (Thiele and Palsson, 2010a) Reactions and metabolites incorporated in each model are presented in **Supplementary Data File 1**. Briefly, the initial version of the network was first reconstructed using the ModelSeed platform (Devoid et al., 2013). Then, the metabolic network was refined based on the information present in the KEGG (Kanehisa et al., 2012) and MetaCyc (Caspi et al., 2008) databases. Genome annotation data was used to confirm gene-to-protein relationships and predicted functions for each gene incorporated in the network. Transporters were identified using the Transporter Classification Database (TCDB; www.tcdb.org (Saier et al., 2014)) and the program GBlast (Reddy and Saier, 2012) using parameters described previously (Buyuktimkin and Saier, 2015). The biomass composition was determined

based on a generic gram-negative composition. The amino acid and lipid compositions were tweaked based on Leptospira-specific literature where available.

4.5.2 Simulations and conversion to a mathematical model

Each model exists as an SBML file. These files were used to perform simulations and constraint-based analyses using COBRApy (Ebrahim et al., 2013) and the GUROBI solver v6.0 (Gurobi Optimization, 2015). The constraint-based model consists of an **S** matrix with x rows and y columns, where x is the number of distinct metabolites (in all three compartments) and y is the number of reactions including exchange and biomass reactions. Each of the reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 mmol $gDW^{-1}\ h^{-1}$ and a lower bound of −1000 mmol $gDW^{-1}\ h^{-1}$, making them practically unconstrained, while irreversible reactions have a lower bound of zero. The GapFind MILP algorithm (Satish Kumar et al., 2007) was used to identify required minimal media components. Metabolic pathways were constructed using Escher (King et al., 2015a). All four models are available as SBML downloads in the BiGG database (http://bigg.ucsd.edu) (King et al., 2015c).

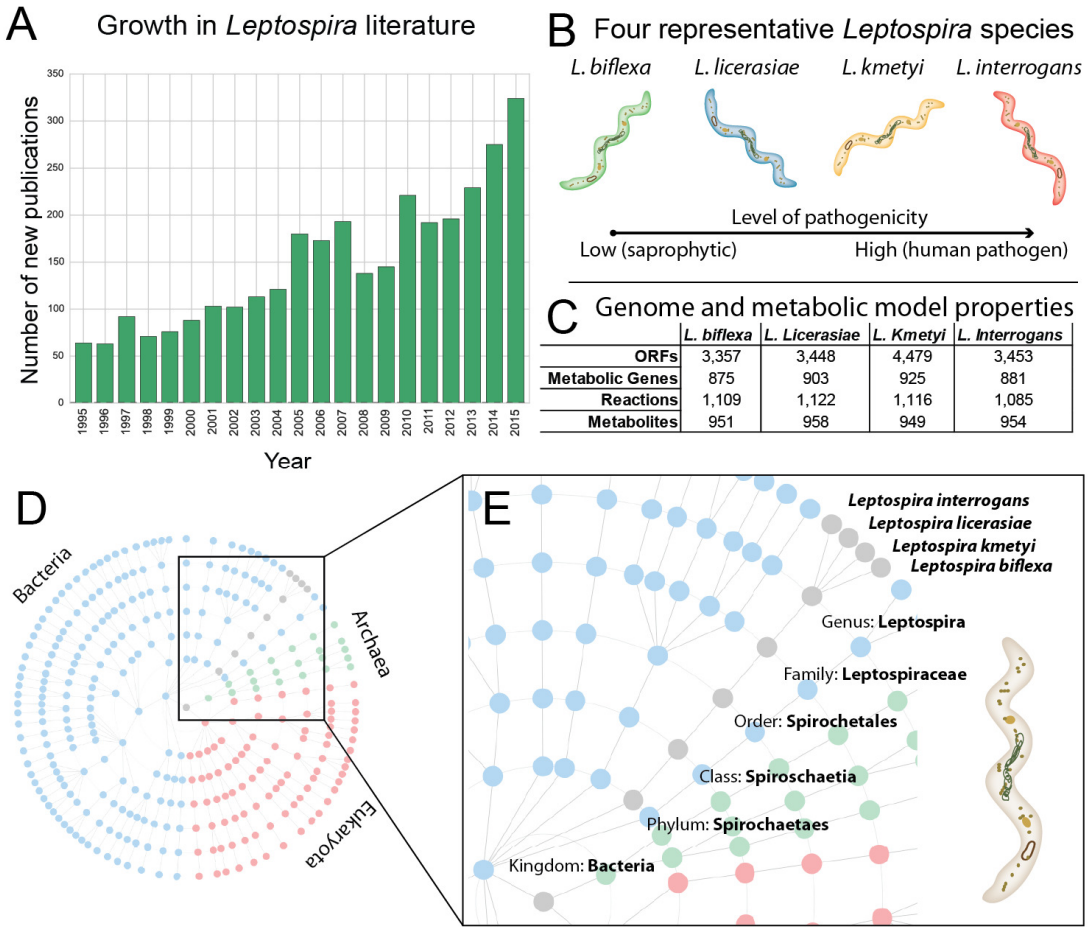
4.5.3 Analysis of essential reactions and metabolites

To simulate the effects of gene knockouts, each model was constrained using default values and the minimal media defined in this study. All reactions in the model were knocked out one a time and growth was simulated by FBA. Reaction knockout strains with a growth rate above zero were considered non-essential. Similar to reaction essential, metabolite essentiality was simulated by removing each metabolite one by one. This was accomplished by deleting all the outgoing (consuming) reactions around the metabolite to be removed. If the removal of a certain metabolite led to zero predicted

cell growth, the metabolite was deemed essential. The essential metabolites were grouped into their relevant metabolic subsytems by taking the largest set of subsytems related to the reactions deleted.

4.6 Acknowledgements

Chapter 4, in part, is a reprint of the material Monk JM**,** Tsu B*, Buyuktimkin B, Saier MH, Palsson BO. Comparative metabolic network analysis and modelling of four Leptospira species provides insight into pathogenesis of Leptospirosis. *In Preparation.* The disseration author was the primary author (equally contributing with Brian Tsu) of this paper.



	L. biflexa	*L. Licerasiae*	*L. Kmetyi*	*L. Interrogans*
ORFs	3,357	3,448	4,479	3,453
Metabolic Genes	875	903	925	881
Reactions	1,109	1,122	1,116	1,085
Metabolites	951	958	949	954

Figure 4.1 Increase in Leptospira knowledge.
Growth in leptospiral literature on NCBI pubmed over the last 20 years. Bar values represent the number of new articles published per year on the topic of Leptospira. B) Summary of four representative species of Leptospira selected for this study. The strains range from non-pathogenic (saprophytic, L. biflexa, to intermediately pathogenic, L. licerasiae and L. kmetyi, to pathogenic, L. interrogans). C) Leptospira species genome and metabolic reconstruction properties. Number of open reading frames (ORFS), metabolic enzyme encoding genes, reactions and unique metabolites. D) Complete and curated metabolic network reconstructions across the phylogenetic tree of life (http://sbrg.ucsd.edu/optimizing-genres) Bacteria (blue), archaea (green) and eukaryota (red). E) Leptospira (grey) are in the bacterial kingdom in the spirochaetes phylum. The reconstructions presented here represent the first curated reconstructions of any species in the spirochaetes phylum and thus their content can be compared to other spirochaetes to aid future reconstruction efforts in this underrepresented phyla.

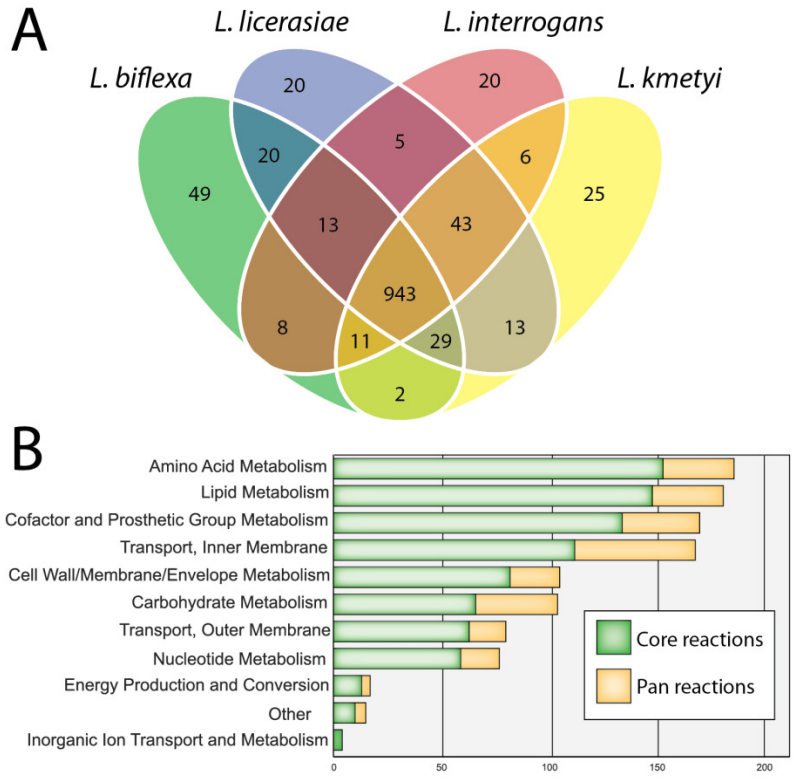


Figure 4.2. Comparison of reaction content among the four leptospira reconstructions. **A)** Venn diagram with shared reaction content and sets of reactions specific to unique models. **B)** Reaction content shared among all four reconstructions is noted as core (green), reactions that are missing in at least one reconstruction are denoted as pan (orange). The reactions are organized by metabolic subsystem.

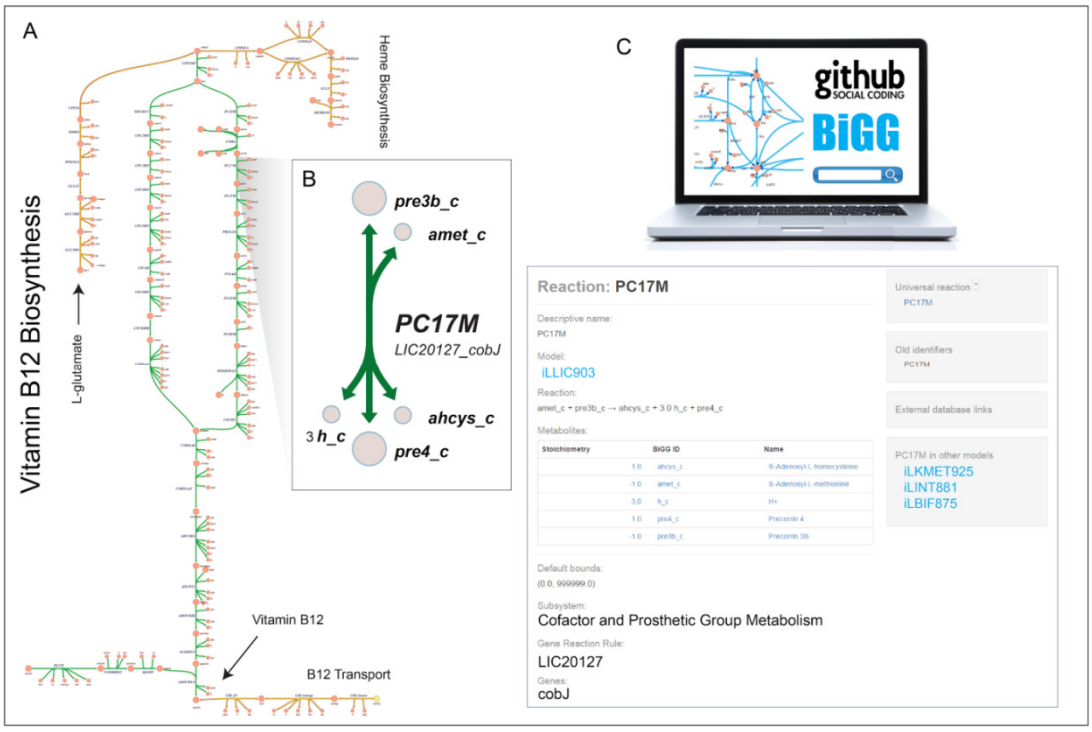


Figure 4.3. Metabolic map of Leptospira interrogans vitamin B12 biosynthesis.
A) The b12 biosynthesis pathway is displayed in detail. Reactions shared by all four leptospires are marked in orange. Reactions specific to individual species are marked in green. **B)** A zoomed-in view of one reaction in the B12 biosynthesis pathway: precorrin-3B C17-methyltransferase (PC17M, EC: 2.1.1.131) along with its catalyzing gene (cobJ) and the locus encoding this gene in L. interrogans. **C)** All metabolic content and maps are available for download and exploration in the BiGG database (http://bigg.ucsd.edu) where all metabolic content can be compared between the species analyzed here and any other organism for which a genome-scale metabolic model is available.

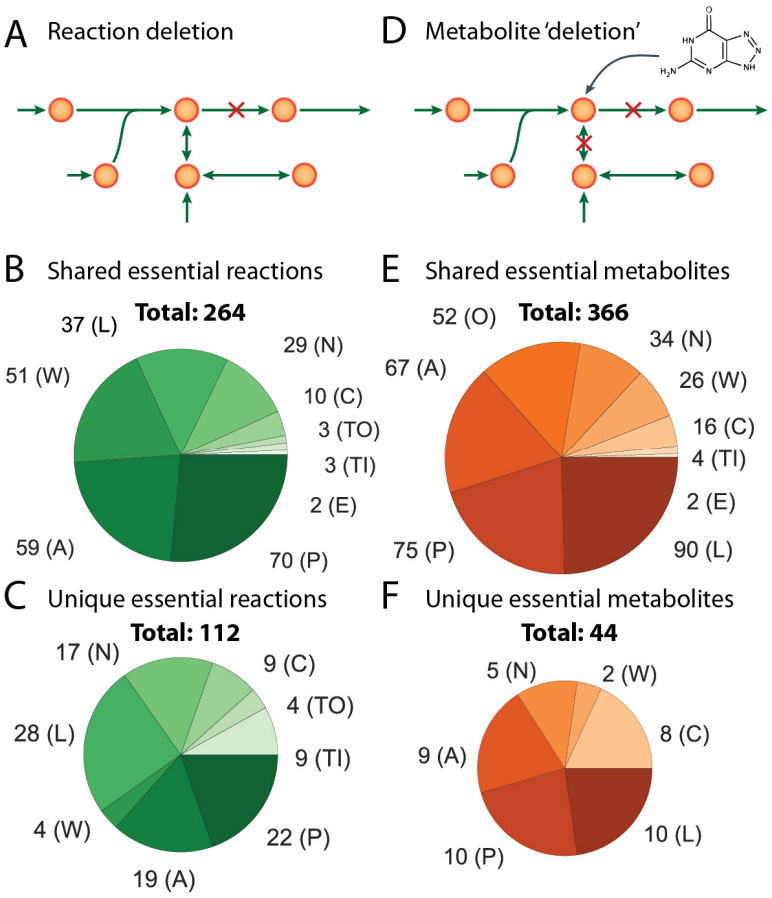


Figure 4.4. Essential reactions and metabolites predicted by genome-scale modelling. **A)** Genome scale models can be used to predict essential reactions (green). All essential reactions were predicted across the four *Leptospira* species of which 264 were universally essential across all species **(B)** and 112 were unique to individual sets of species **(C). (D)** Genome scale models can also predict essential metabolites (orange) by computationally removing consuming reactions of each metabolite in the network. All essential metabolites were predicted across the four *Leptospira* species of which 366 were universally essential across all species **(E)** and 44 were unique to individual sets of species **(F)**. The essential reactions were group by metabolic subsystem. Abbreviations: P: Cofactor and Prosthetic Group Metabolism, A: Amino Acid Metabolism, W: Cell Wall/Membrane/Envelope Metabolism, L: Lipid Metabolism, N: Nucleotide Metabolism, C: Carbohydrate Metabolism, E: Energy Production and Conversion, TO: Transport, Outer Membrane, TI: Transport, Inner Membrane and O: Other.

Chapter 5 Quantifying variation within the bacterial species *E. coli* and the impact on host strain selection

5.1 Abstract

Escherichia coli strains are widely used in academic and applied research as well as in biotechnology for production of various compounds. Despite its status as a model organism, strain-specific differences and their underlying contributing factors are still not well characterized. These differences have a major impact on cell physiology and for the applied purposes of synthetic biology, metabolic engineering, and process design. In this study, strain-specific differences are quantified in seven widely-used applied biotechnology strains of *E. coli* (BL21, C, Crooks, DH5a, K-12 MG1655, K-12 W3110, and W) using genomics, phenomics, transcriptomics and genome-scale modelling to guide the choice of strain for a given product. Even given the genetic similarity of the strains, metabolic physiology and gene expression varied widely with downstream implications for productivity, product yield, and titre. Further, these differences can be linked to differential regulatory structure. Analysing high flux reactions and the expression levels of their encoding genes revealed a quantitative link between these sets and show that often, these sets are correlated with strain-specific caveats. Integrated modelling also revealed that certain strains are better suited to produce a given compound or express a desired construct considering native expression states of pathways that enable high-production phenotypes. The result of this study is a resource comparing strains in an important model species and a general strategy for choosing a host strain or chassis selection for applied biotechnology.

5.2 Introduction

Escherichia coli dominate the world of biological sciences as a model prokaryote for physiology studies, as an important pathogen, and as a key host for metabolic engineering and synthetic biology. This diversity in lifestyle and application reflects the

high level of genetic diversity within the species. Thanks to the genomics revolution in microbiology that has enabled sequencing of diverse strains for any species, it is now know that the genomes of different strains of *E. coli* range in size from 4.5 to over 5.5 Mbp, and the species has a pan-genome composed of more than 15,000 unique proteins (Lukjancenko et al., 2010, Gordienko et al., 2013). Part of this large pan-genome consists of unique metabolic capabilities that have been shown to have important implications for infectious disease studies and pathogenic niches(Monk et al., 2013, Baumler et al., 2011, Vieira et al., 2011). This metabolic diversity is likely to be equally impactful on synthetic biology applications. The massive genomic diversity of the *E. coli* species provides a deep pool of strains to use for basic research and for potential host strains to be chosen for metabolic engineering and synthetic biology applications. It also raises an important question: what range of phenotypic behaviours exist and how can these be leveraged to further exploit *E. coli* as a model organism and host strain?

A review of industrial biotechnology publications and patents that use *E. coli* as a host strain yielded seven representative *E. coli* strains that are used often and are good candidates for detailed study: the K-12 strains MG1655, W3110, and DH5a, as well as strains BL21, C, Crooks and W (**Figure 5.1A**). The selection of both closely related strains (K-12 strains) and more distantly related strains also allowed an examination of whether close genetic relatedness is a useful predictor of physiological relatedness and production potential. The existing body of work evaluating different *E. coli* strains in metabolic engineering and synthetic biology(Archer et al., 2011, Arifin et al., 2014, Yoon et al., 2012, Vijayendran et al., 2007, Marisch et al., 2013, Chae et al., 2010) demonstrated a need for the comprehensive analysis of strain-specific differences. Despite significant success in engineering *E. coli* for industrial production of chemicals

and proteins(Lee et al., 2012b, Kim et al., 2015), there is no unified fundamental basis for selection of one strain over another for a given metabolic engineering project or expression of a given construct. Previous studies have shown that the choice of host strain for production of a given compound has a significant impact on results(Na et al., 2013, Kim et al., 2014) and up until now represented a major brute force screening effort. Thus, an important question remains to be addressed: what strain of *E. coli* is best suited for production of a desired product?

Here, a comprehensive comparison incorporating transcriptomics, genomics, and phenomics with genome-scale modelling of seven common *E. coli* production strains is presented and a mechanistic basis for the selection of a given *E. coli* strain for production of particular compound is established. The data and models are further used to develop a general strategy for synthetic biology host strain selection that can be applied to any production organism with sufficient genetic diversity. The work presented here establishes a workflow and represents a resource for similar efforts with other organisms and/or additional omics data types.

5.3 Results

5.3.1 Whole-genome sequencing and comparative analysis

Seven strains of *E. coli* were sequenced to comprehensively compare and examine their strain-specific genetic differences. Accurate genome sequences were determined to be essential due to recent studies that demonstrate several differences between the reference sequence of *E. coli* K-12 MG1655 and the stock strains of laboratory *E. coli* available from culture collections (Freddolino et al., 2012). These differences were shown to have substantial physiological effects that could confound experimental results and have downstream impacts on bioprocess design(Nahku et al.,

2011). One of the widely-used *E. coli* strains, C, had no public genome sequence available, thus whole genome sequencing was performed to establish the genetic parts list for this strain (see **Methods**). The *E. coli* C draft genome was predicted to be 4.54 Mbp in size and has 4,424 open reading frames.

The whole genome sequences of the seven strains were then used to classify the strains based on their genetic content. First, a classical MLST scheme (Jaureguy et al., 2008) was used to assign the *E. coli* strains to phylogroups. All strains were assigned to group A, containing primarily safe, commensal strains, except for *E. coli* W that was assigned to group B1, a group that contains several pathogenic members. A full genome alignment and comparison of conserved proteins was also performed (**Methods**). A total of 6,626 unique protein-coding sequences were discovered across all seven genomes. Of these, 3,316 genes were shared between all seven strains, forming a “core” genome. Of the non-core genes, 1,493 were present in 2-6 of the strains and 1,817 of the genes were unique to a single strain alone (**Figure 5.1B**). A full-genome DNA alignment showed that the *E. coli* K-12 strains, MG1655, W3110, and DH5a were all part of the same clade. *E. coli* BL21 and C were also part of a similar clade, and *E. coli* Crooks and W strains were separate from the others with *E. coli* W being the most distantly related strain (**Figure 5**/**1C**).

5.3.2 Phenotypic characterization of host strains highlights physiological differences

To assess growth dynamics and by-product secretion rates, phenotypic characterizations were performed in aerobic and anaerobic M9 minimal media (**Methods**, **Supplementary Figure 4**)**.** Major differences were observed between the strains during exponential growth phase. Aerobically, the growth rates ranged from 0.61

h^{-1} (W3110) to 0.96 h^{-1} (W and Crooks), with a mean growth rate of 0.80 ± 0.12 h^{-1}, see **Table 1**. Anaerobically, DH5a grew slowest (0.18 h^{-1}) and W grew fastest (0.90 h^{-1}), with a mean growth rate of 0.53 ± 0.25 h^{-1}. This difference is stark given that the strains share more than 95% of genes in central metabolism at greater than 95% amino acid identity (see **Methods**) indicating vastly different utilization of similar central metabolic genetic content.

While the overall biomass and by-product yields between strains were similar, the strains exhibited different organic acid secretion profiles. In aerobic conditions, four of the strains, C, DH5a, MG1655, and W3110 exhibited acetate overflow metabolism in this well-aerated experiment (**Figure 5.1D**) in agreement with past studies (Archer et al., 2011, Marisch et al., 2013). Anaerobically, all strains exhibited common mixed acid fermentation with production of acetate, formate, ethanol, and succinate. Only two strains, BL21(DE3) and DH5a, produced lactate anaerobically (**Figure 5.1E**). This physiological characterization clearly shows that strains differ in their propensity to make certain molecules, e.g., lactate, an industrially-relevant biologically-produced chemical, when growing in their native state (Jang et al., 2012).

The rate of substrate consumption in the different strains (**Table 1, Supplementary Figure 4)** also exhibited significant variation (a 1.9 and 3.6 fold difference aerobically and anaerobically, respectively), a fact that has important implications for productivity and bioprocessing costs.

5.3.4 Strain-specific genome-scale models (GEMs) of metabolism reveal differences in metabolic capabilities

The large physiological differences across the selected *E. coli* strains motivated the construction of seven strain-specific GEMs (**Supplementary Data Files 1 and 2**)

that were used to integrate, model, and contextualize the measured physiological data. The models were first validated by demonstrating that they could recapitulate a functional flux state by setting the measured physiological data (i.e., inputs and outputs – glucose uptake rate, growth rate, byproduct production rates). All models passed this test, indicating consistency between the models and physiological data (**Supplementary Figure 5**). Next, each model's metabolic content was compared to classify reactions as part of "core" or "pan" metabolic capabilities (**Supplementary Table 2 and 3**). The core content (reactions present in all seven strains) consisted of 1,265 genes, catalysing 2,315 reactions that utilize 1,776 different metabolites. The total content, present in at least one strain, but not shared among all, consisted of 2,526 reactions – indicating that 211 reactions were variably present in different strains. The average model had 2,425 +/- 17 reactions. In a recent study of 55 strains of *E. coli* (Monk et al., 2013) including pathogens and environmental isolates, the average model had 2,337 +/- 52 reactions, indicating that there was more diverse metabolic content among the 55 strains than exists between the seven industrially useful strains examined here. However, several of the differences between the seven strains are present in subsystems important for metabolic engineering, including the pentose phosphate pathway and amino acid biosynthesis. For this reason, strain-specific GEMs of metabolism were used to examine maximum theoretical yield of growth precursors and industrial chemicals to explore the functional differences and metabolic capabilities of each strain.

5.3.5 Strain-specific metabolic models highlight differences in theoretical yields of industrially-relevant compounds

The theoretical yields of industrially-relevant native and non-native compounds were examined by utilizing strain-specific models. A total of 245 heterologous pathways

for the production of non-native compounds from a recent study (Campodonico et al., 2014) were integrated with each strain-specific model to compare theoretical yields. The yields were calculated using glucose as the sole carbon source in both aerobic and anaerobic conditions (**Supplementary Data File 3)**. Overall, the majority of the maximum theoretical yields were similar across strains (83% of pathways had identical maximum yields across the seven strains). However, several differences were identified between the seven strains. For example, the model of *E. coli* BL21 is unable to produce acrylic acid from a specific heterologous pathway (pathway 23) because it lacks N-acetylglucosamine kinase. This means that it cannot make N-Acetyl-D-glucosamine 6-phosphate from N-acetyl D-glucosamine (acgam) – a requirement for this heterologous pathway to produce acrylic acid. Also, DH5a cannot make 3-hydroxypentanoic acid via a predicted heterologous route (pathway 223) due to the lack of homocysteine S-methyltransferase encoded for by *mmuM* (Song et al., 2015). A histogram of differential yield by pathway in each strain is given in **Supplementary Figure 6**.

While most strains have equal theoretical yields, some of the heterologous pathways displayed strain-specific differences (591 of the 3,430 total pathway (245), strain (7) and condition (2) combinations). In this variable set, *E. coli* W and Crooks often had the greatest yield for a given product (max yield in 194 and 102 pathways, respectively, out of 245 pathways in aerobic and anaerobic conditions), while strains C and BL21 (max yield in 40 out of 245 pathways in aerobic and anaerobic conditions) often had decreased yields. Strain BL21 had reduced yields for production of all compounds in aerobic conditions due to the lack of 6-phosphogluconolactonase (PGL) reaction activity (Meier et al., 2012) in the oxidative pentose phosphate pathway (PPP), encoded by the gene *pgl*. This requires an alternate pathway for production of ribulose-

5-phosphate that does not generate NADPH, one of the primary purposes of the oxidative PPP (Fan et al., 2014) (**Supplementary Figure 7**).

Analysis using strain-specific models revealed several increased maximum theoretical yield advantages. *E. coli* Crooks and W had a 4-12% greater yield of 2-oxobutanoate on five of the different heterologous pathways in anaerobic conditions because of an alternate isoleucine biosynthesis pathway (see **Supplementary Text**). Furthermore, models of BL21 and Crooks had 21% higher yield of 1,4-butanediol in anaerobic conditions for two of the heterologous pathways (i.e., pathways 176 and 177) due to the ornithine aminotransferase reaction (see **Supplementary Text**). These differences in maximum theoretical yields demonstrate that major differences in strain behavior exist based solely on internal reaction content and the unique metabolic network structure of each strain. Next, to gain a deeper understanding of strain specific behavior, the measured physiological data was integrated with each strain-specific model.

5.3.6 Integration of phenomics with strain specific models classifies shared and strain-specific high flux pathways

The analysis of theoretical yields presented above represents the maximum (i.e., ideal) capabilities of each strain. *In vivo* wild-type strain-specific behaviour can be analysed by integrating the measured strain-specific physiological data with its corresponding model. The constraint-based modelling techniques of flux variability analysis (FVA) (Mahadevan and Schilling, 2003a) and Monte Carlo Markov Chain (MCMC) sampling (Schellenberger and Palsson, 2009b) were performed to determine minimum, maximum, and likely flux through each reaction in each strain based on the imposed physiological constraints (for example *E. coli* C, **Figure 2A, Supplementary**

Figure 8, Supplementary Data File 5). The resulting probable flux distributions were used to classify reactions that must carry high flux (**Methods**) to achieve the measured physiological secretion and growth rates, and were compared in both aerobic (**Figure 5.2B**) and anaerobic (**Figure 5.2C**) conditions.

High flux reactions were compared across the different strains (**Figure 5.2D**). Aerobically, there were 62 reactions classified as high flux in at least one strain. Of these, 37 were shared among all seven strains. Most of the shared reactions were involved in glycolysis, the TCA cycle, and the PPP (**Supplementary Data File 4**). In addition, reactions involved in glutamate metabolism were classified as high flux across all seven strains. The remaining 25 reactions were classified as high flux in at least one strain, but not shared by all. Some of these differences were obvious on a genetic level – for instance, five reactions in the oxidative PPP were classified as high flux in all strains except BL21, because, as discussed above, BL21 lacks the *pgl* gene, disabling flux through the oxidative PPP in this strain. Other differences in high flux reactions were related to differences in physiological behaviour. For example, acetaldehyde dehydrogenase was only a high flux reaction in two strains (DH5a and MG1655 – two of the strains that exhibited acetate overflow metabolism). Acetate secretion negatively correlated with flux through TCA cycle reactions, including citrate synthase (CS), aconitase (ACONTa/b), and isocitrate dehydrogenase (ICDHyr) (**Supplementary Data File 7, Supplementary Figure 9**)**.** Under anaerobic conditions, there were a total of 64 high flux reactions classified in at least one strain. Of these, 29 reactions shared high flux across all seven strains. These included predominantly glycolysis reactions and pentose phosphate pathway reactions as well as pyruvate formate lyase (PFL).

5.3.7 Transcriptome analysis classifies shared and strain-specific gene expression profiles

To delve deeper into strain-specific behaviour and the observed genetic and physiological differences, RNA-seq was used to collect genome-wide transcriptomic profiles of each strain at exponential phase in aerobic and anaerobic conditions (**Supplementary Figure 10**, **Supplementary Data File 8**). Pairwise differential expression was compared between each of the seven strains (**Supplementary Figures 11 and 12, Supplementary Data Files 9 and 10**) and correlation coefficients were calculated to quantify the level of similarity between full expression profiles of shared genes for the different strains (**Figure 5.3A and B**). A PCA was also performed that focused on metabolic genes (**Figure 5.3C and D**). The analysis highlights major differences in expression states. For example, BL21 displayed significantly different expression profiles in anaerobic conditions due to high expression of TCA cycle genes. This difference is most likely due to a nonsense mutation in the gene encoding the global oxygen-responsive transcriptional regulator FNR (Pinske et al., 2011) making this strain's gene expression behave more similarly to an aerobic state. Further differences are discussed in the **Supplementary Text**.

As with reaction flux, gene expression values were analysed for each growth condition and classified into highly expressed gene sets (**Methods**). This analysis identified a group of genes that were highly expressed species-wide. In aerobic conditions, 199 metabolic genes were classified as highly expressed in at least one of the seven strains (**Supplementary Figure 13**, **Supplementary Data File 11**), but only 11 of these genes were significantly highly expressed across all strains. Three of these were involved in glycolysis: enolase (*eno*), fructose-bisphosphate aldolase (*fbaA)*, and glyceraldehyde-3-phosphate dehydrogenase (*gapA*). In anaerobic conditions, 174

metabolic genes were classified as highly expressed in at least one of the strains, and 23 of the genes were highly expressed in all seven strains including *eno* and *fbaA* as well as acetaldehyde dehydrogenase (*adhE*) and methionine adenosyltransferase (*metK)*.

5.3.8 Transcription factors involved in differential regulation illuminate differential regulatory strategies

The major differences observed in transcription profiles demonstrate unique regulatory mechanisms between strains. Knowledge of transcriptional control is directly applicable to bioprocessing and synthetic biology applications for tuning gene expression levels. Most transcription factors (TFs) have been characterized in *E. coli* K-12 MG1655, thus gene expression profiles between this strain and the other six were compared in both aerobic and anaerobic conditions. An enrichment analysis of TFs known to regulate gene expression was performed (see **Methods**). There are 196 TFs with known regulons available in Regulon DB (Huerta et al., 1998). For each strain, an average of 28±3 TFs were enriched for differential control of expressed genes in aerobic conditions and 29±6 TFs were enriched in anaerobic conditions (**Supplementary Data File 12**). An informative example is that of the galactitol regulon which includes *gatYZABCD* and is negatively repressed by the *gatR* TF (Nobelmann and Lengeler, 1995). The *gatR* TF is highly enriched for differential expression in all of the strains except W3110. In MG1655 and W3110 the *gatR* gene has an IS3E insertion leading to constitutive expression of these genes (Nobelmann and Lengeler, 1996). This aberrant regulation leads to expression and translation of *gat* genes that are ultimately responsible for nearly 1% of the wild-type *E. coli* K-12 MG1655 proteome (Li et al., 2014). In the other strains, *gat* gene expression is low due to repression by *gatR*.

Other TFs that were significantly enriched for differential expression include, in aerobic conditions: *arcA* (anoxic redox control), *cra* (the catabolite reporessor activator), and *gadE* (glutamic acid decarboxylase involved in maintenance of pH homeostasis), and anaerobically: *fnr* (mediates aerobic to anaerobic transistion), iHF (integration host factor, responsible for maintaining DNA architecture), and *purR* (controls purine nucleotide biosynthesis) (**Supplementary Table 4**). Examining TF enrichment between strains identifies unique, strain-specific control mechanisms for different genes, even those that are conserved between strains. Further analysis will aid in determining differential regulatory mechanisms between strains of *E. coli* with the ultimate goal of manipulating gene expression to enhance metabolic engineering strategies.

5.3.9 Intersection of high flux pathways with highly expressed genes

A quantified correlation between high flux reactions and gene expression is key to understanding overall cell physiology and is of great interest to industrial biotechnology as overexpression of genes desired to carry high flux is a widely-adapted approach to increase production of a target molecule (Lee et al., 2012a). In this study, 50±8% of model-determined high flux reactions also had encoding genes that were highly expressed. This overlap occurred significantly more often than random (empirical p-value < 0.001, permutation test, see **Methods, Supplementary Figure 14**). Several genes, such as *eno, fbaA,* and *gapA,* were consistently high flux and highly expressed in all seven strains (**Figure 5.4A**, **Supplementary Data File 13**). Other gene/reaction pairs were less conserved, including those involved in amino acid metabolism such as *ilvD, serC, and aspC*, perhaps indicating large differences in amino acid use and biosynthesis between each of the strains. While a correlation between high flux reactions and gene expression is observed, it is unsurprising that several genes/reactions do not correlate

as recent work has demonstrated that gene expression can be a poor indicator of enzymatic activity (Machado and Herrgard, 2014).

Prior to determining which strain might be best suited to produce a given target compound, an analysis was performed to answer the question of whether GEMs can be used to *a priori* predict changes in gene expression from one state to another. Using physiological data in aerobic and anaerobic conditions, the fluxes were predicted for a shift from aerobic to anaerobic conditions (**Supplementary Figure 15**). Overlap between model-predicted changes in reaction fluxes and experimentally observed changes in gene expression were analysed. On average, the metabolic models correctly predicted major changes in flux during a shift from aerobic to anaerobic conditions for 82±8% of the major reaction flux changes (30±12 genes per strain, see **Supplementary Table 5 and 6, Supplementary Data File 14**). The results of this analysis indicated a level of predictability suitable for *de novo* strain-specific prediction in production strains (examples are given in **Figure 5.4B-C**).

5.3.10 Model-driven analysis of production potential

An analysis was performed to determine the strain best suited for the production of a given compound as well as expression of a given construct from the set of *E. coli* strains examined in this study. A common metabolic engineering approach is to increase expression of the genes in a pathway of interest that lead to a product (Lee et al., 2007, Lee et al., 2012a, Huo et al., 2011). Based on this approach, it was reasoned that strains with natively high expression in a pathway of interest are likely better poised to produce a given product, as they would require less interventions to achieve a production goal. Therefore, genome-scale modelling was integrated with expression data to determine strains that are inherently best poised for production of a given product. Strain-specific

models were used to predict the optimal flux distribution for production of two different sets of compounds in aerobic and anaerobic conditions: 1) all 20 amino acids using native *E. coli* pathways and 2) 20 non-native compounds using 245 heterologous pathways (Campodonico et al., 2014) (**Supplementary Data File 15**). Combining predicted fluxes with gene expression values allowed for the generation of a relative production potential score ('R-score', see **Methods**) that gauges a strain's suitability for producing a given compound (e.g., **Figure 5.5C** and **5D**).

An integrated analysis using transcriptomic data and genome-scale modelling revealed that each of the seven strains may be preferentially suited for production of different target metabolites. Strains that most often had an R-score >1 for amino acid production were MG1655 and DH5a for aerobic conditions (12/20 and 5/20, respectively) and MG1655 and W for anaerobic conditions (7/20 and 3/20, respectively). The targeted product also highlighted strain-specific differences. For example, in aerobic amino acid over-production (**Figure 5.5A**), it was found that *E. coli* W was predicted to be better at production of pyruvate-derived amino acids leucine and valine due to a more than two-fold greater expression of *leuC*, *leuD,* and *ilvE* compared to the other six strains. Variations in production potential were also prevalent across the 245 heterologous pathways examined (corresponding to one of 20 different industrial compounds, some targeted products originated from multiple native precursors in the cell). Similarly, R-scores >1 were distributed across all seven strains examined. K-12 MG1655 had the highest number of R-scores >1 for 94 pathways aerobically, and W and C had 42 and 41 under anaerobic conditions, respectively (**Figure 5.5B**).

Grouping the 20 different targeted heterologous products leads to a further characterization based on which strains were best suited for production of a particular

class of compound. For example, strain W was best suited for production of 5/20 compounds (2-methyl-1-butanol, 1-butanol, 3-methyl-1-butanol, 2-keto-isovaleric acid, and 2,3-butanediol) independent of the heterologous pathway used (**Supplementary Table S7**). In contrast, the best production strain for 1,4-butanediol varied based on the heterologous pathway used. For example, strain K-12 MG6155 had high expression of 2-oxogluterate dehydrogenase encoding genes *sucA, sucB, and lpd* (2-fold greater than expression for strains C, Crooks, DHa, and W3110) that produce succinyl-CoA, a branch point for several of the pathways leading to 1,4-butanediol production. However, other heterologous pathways leading to production of 1,4-butanediol start from 4-aminobutanal and DH5a was predicted to be best suited for these pathways.

Extending the model-driven analysis to selection of host strains (i.e., chassis) for synthetic biology applications revealed strain preferences based on amino acid requirements of a given construct. Coding sequences of synthetic biology constructs were obtained from the registry of standard biological parts and their amino acid composition were calculated. Further, the overall amino acid makeup of the *E. coli* proteome is stable (Li et al., 2014) and this trend holds true for amino acid frequencies across bacteria (Gilis et al., 2001, Hormoz, 2013, Latif et al., 2015). Thus, constructs with amino acid compositions that are significantly over-represented may require higher demand for a given amino acid if the goal is to significantly produce the contract as a large part of the host strain's proteome. Analysing this concept, the R-score analysis for amino acid production capability was applied to each construct by comparing the overlap of a strain's highly expressed amino acid biosynthetic pathways (found to be 1-4 amino acid pathways per strain based on the R-score) with those overrepresented in each construct. This approach led to a prediction of which strains may be best at expressing a

certain synthetic biology construct considering both construct required and total amino acid pathways enriched in a strain (**Supplementary Figure 16**). Under aerobic conditions, strain DH5a was predicted to be the best producer for the most constructs (568/3,983 or 14% of constructs) due to its inherent high expression of the biosynthetic pathways for tyrosine (Y) and phenylalanine (F) (amino acids that are often small fractions of the proteome, **Supplementary Figure 16A**) followed by BL21 (473/3,983 or 12% constructs) for similar reasons. This result aligns well with the fact that DH5a is often preferred and used in cloning applications (Taylor et al., 1993, Song et al., 2015) and BL21 is popular for expression of recombinant proteins (Robichon et al., 2011, Marisch et al., 2013).

In summary, this approach emphasized the importance of strain-specific advantages in terms of network structure and native expression states that should be considered when choosing a host strain or chassis. Full results are provided in **Supplementary Data Files 15 and 16**.

5.4 Discussion

This study establishes a workflow to compare and presents a resource on the important bacterial species *E. coli* that was used to guide selection of the best host for applied biotechnology. The omics data generated here addresses a gap in *E. coli* knowledge comparing this well-known species and in strain-specific information for seven industrially important strains grown in two well-defined conditions. This unified multi-omics dataset was integrated with GEMs to characterize strain-specific and species-wide properties of *E. coli* by comparing metabolic fluxes, gene expression, and differential regulation across the strains. New, quantified relationships between these datasets were drawn, along with an evaluation of the production potential of the strains

based on maximum theoretical production yields and strain-specific native expression states. The compendium of data, GEMs, and production pathway analyses presented here provide the basis for analysing the overall diversity and production capabilities of the seven *E. coli* strains studied. Key findings are available in **Supplementary Table 8**.

A number of important strain-specific and species-wide properties for *E. coli* were identified. The K-12 strains are genetically very similar considering the overall genetic diversity of the 7 strains, yet their expression profiles under aerobic conditions showed significant variability (**Table 5.1**). Previous studies have shown that W3110 has an amber mutation (stop codon) at position 33 in *rpoS* which is not found in MG1655 (Vijayendran et al., 2007). This mutation has been shown to reduce RpoS activity (Subbarayan and Sarkar, 2004). RpoS is one of the primary global regulators of *E. coli's* complex regulatory network. Thus, a small change can have a large effect on cellular expression patterns. This highlights the need to better understand and elucidate transcription factor network architecture in even closely related strains of *E. coli;* this data resource enables such a study.

The phenotypic differences observed between the strains, despite the fact that they have largely similar genomes and metabolic reaction networks compared to other sequenced *E. coli* strains (Monk et al., 2013, Baumler et al., 2011, Vieira et al., 2011), were among the most striking results from this study. The glucose uptake rates measured for the different strains were observed to vary more than 3-fold in anaerobic conditions. If the measured wild-type uptake rates can be even partially conserved when generating a bioprocessing strain, selection on this criterion alone could have major implications for strain productivity and bioprocess titres (Arifin et al., 2014). Also, there are a number of cases where some strains have additional or are lacking certain

metabolic enzymes. The maximum theoretical production analysis presented here (**Supplementary Figure S5**) demonstrates that these details are crucial to consider when selecting strains for a metabolic engineering project. Further, the pan-genome of this set is relatively small compared to all *E. coli* strains which have been sequenced thus far (Gordienko et al., 2013), implying that other strains may have pathways and enzymes available to mine for production purposes. Another key result was the identification of a 50±8% overlap of high-flux reactions with highly-expressed genes that is in line with other studies (Holm et al., 2010, Ishii et al., 2007). This significant overlap defines an expected outcome for such data sets. Failure modes may be unnecessarily expressed for a given bioprocess and are therefore targets for expression reduction.

Maximum theoretical production and the native expression state of the cell are important considerations when choosing a strain. The case studies presented here show that specific strains have unique flux and gene expression patterns that, in turn, may affect the production capacity of compound or construct. The native expression of genes within a pathway of interest is not the only factor influencing the generation of a successful production strain. For example, *E. coli* strain DH5a is often used in cloning applications due to an *endA1* mutation that inactivates an intracellular endonuclease (Taylor et al., 1993) and BL21 is well established in recombinant protein production due to a lack of the Lon and OmpT proteases (Ratelade et al., 2009). Thus, aspects such as transformation efficiency (Liu et al., 2014c), phage resistance (Furukawa and Mizushima, 1982), product tolerance (Lennen and Herrgard, 2014) and other traits must also be considered. Furthermore, maximizing theoretical yield does not necessarily lead to increases in titre or productivity. However, the workflow presented here, combining GEMs and omics data, could result in significant time and cost savings by reducing the

number of genetic modifications necessary to develop high-level production strains or find a host to produce a construct of interest in a sufficiently high amount.

The new multi-omics data set provided in this study was generated using consistent and defined conditions for multiple strains of a species. Combined with the integrated analysis performed here, it will be of great use for industrial, basic biology, and human health applications. For example, this data and the R-score method could be applied to examine the production of reactive oxygen species across different strains to determine the impact on antimicrobial treatment (Brynildsen et al., 2013a, Adolfsen and Brynildsen, 2015). This unified and normalized data set allows one to quantitatively compare strains and represents a comprehensive compendium of unique strain characteristics. The generation of similar datasets integrated with genome-scale modelling will enable rational strain-selection and design for metabolic engineering and synthetic biology projects in other common production host organisms.

5.5 Material and Methods

5.5.1 Bacterial strains, media and growth conditions

Escherichia coli strains *E.coli* C (DSMZ 4860), *E. coli* Crooks (DSMZ 1576), *E. coli* DH5α (DSMZ 6897) *E. coli* W (DSMZ 1116), *E. coli* W3110 (DSMZ 5911) were obtained from DSMZ-German Collection of Microorganism and Cell Cultures; *E. coli* BL21 (DE3) was purchased as competent cells from Agilent (Agilent Technologies Inc., USA), *E. coli* K-12 MG1655 (ATCC 700926). All strains were cultured in M9 minimal medium (Miller, 1972) containing Na_2HPO_4 x $7H_2O$ (6.8 g), KH_2PO_4 (3 g), NaCl (0.5 g), NH_4Cl (1 g), $MgSO_4$ (2 mmol), $CaCl_2$ (0.1 mmol), trace elements, Wolf's vitamin solution (Atlas, 2010) and glucose (2 g L^{-1}). Anoxic M9 minimal media with glucose was obtained by flushing solution with oxygen free nitrogen (95%). Overnight cultures from single

colonies of *each of seven E. coli* strains were diluted to a starting optical density (OD_{600}) of 0.01. Cultures were grown in 250 ml flasks or 300 ml oxygen-free sealed bottles containing 50 ml glucose-M9 minimal media in a shaking incubator at 37℃ and 250 rpm.

5.5.2 Growth rates and analytical measurements

Bacterial growth and cell numbers were determined by OD_{600} measurements using an Evolution 220 UV-Visible spectrophotometer (Thermo Fisher Scientific, Germany) with 10 mm optical-path cuvettes. Growth curves of the seven strains were performed in glucose-M9 minimal media. For growth curves, the starter cultures of all strains were pre-adapted to the medium to be tested, for 72 h. Samples were taken in regular intervals and OD_{600} was measured as three independent replicates. The concentration of glucose and organic acids (acetate, formate, lactate, succinate, ethanol) were determined by high-performance liquid chromatography (HPLC), using Ultimate 3000 pump, (Dionex, USA) fitted with an Animex HPX- 87 H column (300 × 7.8 mm) (BioRad, USA) eluted isocratically with 5 mM H_2SO_4 at 30℃ and a flow rate of 0.6 $ml.min^{-1}$. The compounds were detected using UV/VIS (Dionex UVD170U/340U) and refractive index (Shodex RI-101, Japan) detectors. Ethanol production rates had to be estimated for 4 of the 7 strains (C, Crooks, W, W3110) using the measured acetate production rate as a proxy as there was a complication measuring ethanol in some of the samples for these strains (likely due to evaporation). This assumption was justified given that the accurate measurements for ethanol in the 3 other strains resulted in approximately equal acetate and ethanol production rates.

5.5.3 Genomic DNA extraction and DNA sequencing

Genomic DNA was extracted from 5 ml of overnight cultures of seven *E. coli* strains using QIAamp DNA Mini Kit (QIAGEN, Germany). The genomic libraries were generated using the TruSeq DNA Sample Preparation Kit (Illumina Inc., USA). Briefly, 1 µg of bacterial genomic DNA was fragmented for 40s using Crimp Cap microTUBE and Covaris AFA System (Covaris E220, USA). The ends of fragmented DNA were repaired by T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase. The Klenow exo minus enzyme was then used to add an 'A' base to the 3' end of the DNA fragments. After the ligation of the adapters to the ends of the DNA fragments, samples were subjected to 2% 1×TAE agarose gel electrophoresis. DNA fragments ranging from 300 - 400 bp were recovered from the gel and purified using the MinElute Gel Extraction Kit (QIAGEN, Germany). Finally, the adapter-modified DNA fragments were enriched by PCR and normalised to the final concentration of 10 nM. The libraries were sequenced using the Illumina MiSeq platform with a paired-end protocol and read lengths of 150 nt. Differences between reference sequences and those determined were analyzed with breseq. Genome alignments were performed with the MAUVE suite v2.3.1 (Darling et al., 2010). Genome ring images were generated with BRIG v0.95 (Alikhan et al., 2011)

5.5.4 Total RNA extraction and mRNA enrichment

Cells were harvested from aerobic and anaerobic cultures of seven *E. coli* strains grown to an OD600 of 0.6 (exponential phase). Cultures were divided into 10 ml aliquots and were immediately mixed with 0.2 volumes of ice-cold STOP solution (95% ethanol, 5% phenol (pH 4.7)). After 20 min incubation on ice, samples were spun down for 10 min at 4˚C and 7000 x g in a centrifuge. Pellets from one aliquot were gently resuspended in RNAProtect (QIAGEN, Germany) to further stabilize the RNA. Remaining samples were mixed with RNAlater (QIAGEN, Germany) and placed at -80˚C for archival storage. Total

RNA was extracted using RNeasy Mini kit (QIAGEN, Germany) and on column DNase treatment following the manufacturers' instructions. The 23S and 16S rRNAs were removed by subtractive hybridization using the MICROBExpress kit (Ambion, USA) with modifications. Compared with the standard protocol, 50% more capture oligonucleotides and magnetic beads were used. 5S rRNAs (120 nt in length) were removed during the total RNA extraction on column. Specifically, ribosomal depletion on total RNA isolated from the *E. coli* BL21 (DE3) was performed using RiboZero (Gram Negative Bacteria) kit (Epicenter, USA). RNA samples were stored at -80℃.

5.5.5 cDNA library preparation, RNA sequencing and assessment

The sequencing libraries were constructed using the TruSeq RNA Sample Preparation kit (Illumina Inc., USA). Each library was prepared with RNA isolated from seven *E. coli* cultures grown in triplicate to an exponential phase under aerobic and anaerobic conditions. RT-PCR was performed with SuperScript® II One-step RT-PCR reagents (Invitrogen, USA). The libraries were sequenced using the Illumina HiSeq2000 platform with a paired-end protocol and read lengths of 50 nt. The final concentration of DNA and RNA was measured using a Qubit 2.0 Fluorometer (Invitrogen, USA). The integrity of total RNA, DNA contamination, removal of rRNAs and cDNA library validation were assessed with Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

5.5.6 Transcriptome Analysis

Gene expression data (3 biological replicates per strain) were analyzed in the statistical software program R (www.r-project.org) using the *EdgeR* package (Robinson et al., 2010). Data were normalized using the CQN package, which accounts for both gene length and GC content effects (Hansen et al., 2010). Differentially expressed genes were determined by comparing expression values under anaerobic and aerobic

conditions. Those genes with adjusted *P* values less than 0.01 (i.e., false discovery rate less than 1%) were identified as significantly differentially expressed genes. Finally, gene annotations were automatically made using the biomaRt package (Durinck et al., 2005) together with the annotation files available at the Ensembl database (www.ensembl.org) and gene set enrichment analysis (GSEA) was performed using the piano package (Varemo et al., 2013). All R packages used in this study are available in Bioconductor (www.bioconductor.org).

5.5.7 In-silico modelling growth conditions

Each model was exported as an SBML file and used to perform simulations and constraint-based analyses using COBRApy (Ebrahim et al., 2013) and GUROBI linear programming solver. The constraint-based model consists of an **S** matrix with rows representing the number of distinct metabolites (in all three compartments) and columns representing the number of reactions including exchange and biomass reactions. Each of the reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 mmol gDW−1 h−1 and a lower bound of −1000 mmol gDW^{-1} h^{-1}, making them practically unconstrained, while irreversible reactions have a lower bound of zero.

By default, the core biomass reaction is set as the objective to be maximized. Certain reactions are by default constrained to carry zero flux to avoid unrealistic behaviors. These reactions are CAT, DHPTDNR, DHPTDNRN, FHL (formate hydrogen lyase), SPODM, SPODMpp, SUCASPtpp, SUCFUMtpp, SUCMALtpp, and SUCTARTtpp. CAT, SPODM, and SPODMpp are hydrogen peroxide producing and consuming reactions that can carry flux in unrealistic energy generating loops. DHPTDNR and DHPTDNRN form a closed loop that can carry an arbitrarily high flux.

The succinate antiporters SUCASPtpp, SUCFUMtpp, SUCMALtpp, and SUCTARTtpp can form unrealistic flux loops with other transporters for aspartate, fumarate, malate, and tartrate. The genes encoding FHL are known to be active under anaerobic conditions, this reaction is constrained to zero in aerobic conditions to avoid unrealistic aerobic hydrogen production. The non-growth associated maintenance (NGAM) constraint is imposed by a lower bound of 3.15 mmol gDW^{-1} h^{-1} on the reaction ATPM. The exchange reactions that allow for extracellular metabolites to pass in and out of the system are defined such that a positive flux indicates flow out. All exchange reactions have a lower bound of zero except for glucose (–10 mmol gDW^{-1} h^{-1}), the vitamin B12 precursor cob(I)alamin (–0.01 mmol gDW^{-1} h^{-1}), and oxygen and all inorganic ions required by the biomass reaction (–1000 mmol gDW^{-1} h^{-1}). The default lower bound on glucose uptake is based on typical glucose uptake rates. Because only a very small amount of B12 is required for growth, the lower bound on cob(I)alamin uptake is arbitrary and never actually constraining in practice. All models also includes drain reactions for six cytoplasmic metabolites without known consuming reactions that must be drained from the system to allow simulation of steady-state cell growth. These metabolites are *p*-cresol, 5′-deoxyribose, aminoacetaldehyde, s-adenosyl-4-methylthio-2-oxobutanoate, (2r,4s)-2-methyl-2,3,3,4-tetrahydroxytetrahydrofuran, and oxamate.

5.5.8 Markov chain Monte Carlo Sampling procedure

The distribution of feasible fluxes for each reaction in the models used here were determined using Markov chain Monte Carlo (MCMC) sampling (Schellenberger and Palsson, 2009b). We used optGpSampler (Megchelenbrink et al., 2014) to uniformly sample the constrained solution space for each model in both aerobic and anaerobic conditions. The models were first constrained using the measured physiological uptake

rates by adjusting glucose uptake rate, maximum growth rate and by-product secretion rates for acetate, succinate, lactate, formate and ethanol. Using these constraints, FVA was performed to constrain each reaction to its minimum and maximum possible fluxes. To model more realistic growth conditions, sub-optimal growth was modeled. Specifically, the biomass objective function, a proxy for growth rate (Feist and Palsson, 2010b), was provided a lower bound of 90% of the optimal growth rate as computed by flux balance analysis (Orth et al., 2010b). Thus, the sampled flux distributions represented sub-optimal flux-distributions, but still simulated fluxes relevant to cell growth and maintenance. MCMC sampling was then used to obtain thousands of feasible flux distributions (referred to here as "points") using the artificially centered hit-and-run algorithm (Megchelenbrink et al., 2014). For each reaction, a distribution of feasible steady-state flux values was acquired from the uniformly sampled points, subject to the network topology and model constraints. Some reactions were disabled to prevent unrealistic flux distributions. The reactions F6PA and DHAPT are known to be utilized during growth on glycerol but are not active in growth on glucose (Gutknecht et al., 2001). In sampling results these two reactions split flux away from upper glycolysis, thus they were constrained to carry zero flux in all conditions. All sampling was performed with a step size of 100 for 10,000 points.

5.5.9 Yield analysis for production of native and heterologous metabolites

The maximum theoretical yield for native compounds was calculated by creating demand reactions for each metabolite in the model and optimizing for flux through the demand reaction while maintaining 10% of the maximum flux through the biomass objective function calculated using FBA. Yields for the heterologous pathways were calculated in a similar manner after adding the required heterologous pathways from

Campodonico et al. (Campodonico et al., 2014). All maximum theoretical yields were calculated using the standard *E. coli* uptake rates documented previously (Orth et al., 2011). Anaerobic conditions were simulated by setting the maximum oxygen uptake rate to 0.

5.5.10 Classification of high flux reactions and highly expressed genes and correlation coefficients

Reactions were classified as high flux (HFR) if their mean sampled distribution value was greater than 1.5 times the standard deviation over the mean of the absolute value of all mean sampled flux values:

$$HFR = \mu + 1.5 * \sigma l.$$

Genes were classified as highly expressed (HEG) if their expression count value (determined post normalization by CQN was greater than 0.5 times the standard deviation above the mean expression value for all metabolic genes:

$$HEG = \mu + 0.5 * \sigma.$$

Metabolic genes are defined as those that are accounted for in the metabolic model. The Pearson correlation coefficient was calculated between measured physiological data and both gene expression and reaction fluxes for each of the seven strains. Correlations were calculated using python pandas (McKinney, 2011)

5.5.11 Scoring scheme to compare model predicted flux changes with gene expression shifts

A scoring scheme was developed to compare flux and expression shift changes. The scheme awarded model predictions of flux shifts that aligned with changes in gene expression (predicted flux increase, gene expression of encoding genes increased) and penalized predictions that were opposite in direction of gene expression. Reactions were classified as increasing or decreasing in a shift from aerobic conditions of the overlap if

their absolute flux change was greater than 10% of the glucose uptake rate and sampled distributions had minimal overlap. Changes in gene expression were classified as significant if their log2-fold change was greater than half of the standard deviation of the mean expression changes and the their false discovery rate was less than 0.00001.

With this shift set the corresponding catalyzing genes were analysed to determine if those were indeed the ones that changed significantly. A scoring scheme was developed to evaluate how well model predictions aligned with shifts in gene expression. The scheme awarded model predictions of flux shifts that aligned with changes in gene expression (predicted flux increase, gene expression of encoding genes increased) and penalized predictions that were opposite in direction of gene expression. Overall, this analysis led to the conclusion that the models could accurately predict major changes in gene expression.

5.5.12 Analysis of major and minor isozyme transcript ratios

In order to evaluate the correlation between high flux reaction and highly expressed genes, it was first necessary to understand isozyme behavior for targeted reactions as many of them have multiple catalytic isozymes in the *E. coli* GEM of metabolism. This is because the reconstruction process is intended to be comprehensive and therefore lists all possible genes that can catalyze a reaction. For high-flux reactions known to be catalyzed by more than one gene (isozymes) an analysis of the expression level for each catalyzing gene was performed to determine if any one gene was expressed significantly higher across all conditions analyzed, or uniquely to aerobic vs anaerobic conditions. For cases where there was a clear minor isozyme lowly expressed, it was necessary to remove them from the analysis of high flux vs. highly expressed reactions. Isozymes identified this way were compared to a literature search

of enzyme efficiencies that was performed to define major and minor isozymes (Nakahigashi et al., 2009a, Kumar and Maranas, 2009, Covert et al., 2004). Since all of the targeted reactions are well studied and the accepted major or minor roles of most isozymes are known (Sprenger, 1995, Keseler et al., 2009), all minor isozymes were removed from the comparison of high-flux reactions to highly expressed genes. Gene expression transcripts were compared between major and minor isozymes. For the most part, genes classified as major isozymes were expressed at least 2 fold greater than their counterpart minor isozymes in all 7 examined *E. coli* strains (**Supplementary Table S3**).

5.5.13 Transcription factor enrichment analysis

Transcription factors and the genes they are known to regulate were downloaded from regulon dB (Huerta et al., 1998) on 8/10/2014. This list was used to perform hypergeometric enrichment analysis on the differentially expressed genes determined from RNAseq analyses. The scipy (Jnes E et al., 2001) stats package hypergeom was used to calculate hypergeometric enrichment values.

5.5.14 Comparison of core metabolic reaction content

The genes and reactions involved in “core” metabolism were defined as those present in the core metabolic model (Orth et al., 2010a). These include 56 genes, 96 reactions and 72 metabolites involved in subsystems such as glycolysis/gluconeogenesis, pentose phosphate pathway and the TCA cycle.

5.5.15 Heterologous and native pathway scoring

Heterologous pathways were added to each strain-specific model individually. Each model was optimized for a ‘demand’ reaction that consumes the target metabolite

(or exchange reactions for native compounds, e.g. amino acids). The standard uptakes rates were used for these simulations (not strain specific rates). Flux variability analysis was run to determine minimum and maximum fluxes for all reactions when optimizing for the production of a given compound, then the models were sampled as described above. The sampled flux distributions were used to determine active pathways used for production of a given target compound.

The flux through the network was traced backwards from a target metabolite recursively by searching for all reactions that could produce a given compound, then following those reactions backwards to their substrates if they met a flux cutoff of >0.5x max production. Currency metabolites (e.g. atp, nadh, h2o) were filtered from this analysis. The pathways were traced back until they hit reactions classified in the subsystem of 'glycolysis/gluconeogenesis' or until their flux split such that the flux through the reaction was <0.5x max production. This analysis resulted in reactions that were required to carry high-flux for max production of the target compound.

$$\text{P-score} = \sum_{i=1}^{p} \log(|Vi| * Zi)$$

p = pathway length, $|V|$ = absolute reaction flux, Z = reaction catalyzing gene z-score, log of negative numbers were set to 0

$$\text{R-score} = \frac{P\text{-}score - \mu}{\sigma}$$

σ = P-score standard deviation, μ = P-score mean, for a given pathway

Next, we multiplied the flux of each reaction by the measured gene expression z-score determined for the gene that codes the gene product involved in that reaction's catalysis (for isozymes the highest expressed gene's expression was used and for catalytic-complexes, the average of gene expression values was used). We log transformed and summed up the result of this reaction flux combined with gene

expression to establish a final score for each strain's estimated production capacity for a given compound.

To determine outliers in this scoring scheme (strains that would be particularly good or bad at a given compound's production) we calculated the number of standard deviations away from the mean a given score was. All those strains that had a score >1 standard deviation away from the mean production score were predicted to be particularly well suited for production of the target product compared to other strains.

5.5.16 Synthetic biology construct production potentials

All synthetic biology constructs were downloaded from the Registry of Standard Biological Parts. Nucleic acid sequences were filtered for coding regions, yielding 3,982 constructs that were then translated to determine amino acid composition. The average abundance of protein in *E. coli* was taken from two ribosomal profiling datasets (Li et al., 2014, Latif et al., 2015) for *E. coli* growing in different medias. The amino acid composition of each *E. coli* protein was calculated and multiplied by its abundance to approximate the m(Li et al., 2014)(Li et al., 2014)(Li et al., 2014)(Li et al., 2014)(Li et al., 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014)(Li, et al. 2014) lean abundance of each amino acid in *E. coli* (**Supplementary Dataset 16**). Synthetic biology constructs that had amino acid abundances greater than 10 standard deviations above the mean amino acid abundance in *E. coli* were considered to be in demand for production of that given construct. The average construct required 4 amino acids in abundancies greater than the mean with proline (P), tryptophan (W), histidine (H), cysteine (C) and Tryptophan (Y) most often required in greater amounts. Next, R-scores from the amino acid production potential analysis were used to determine each strain's production affinity for amino acid

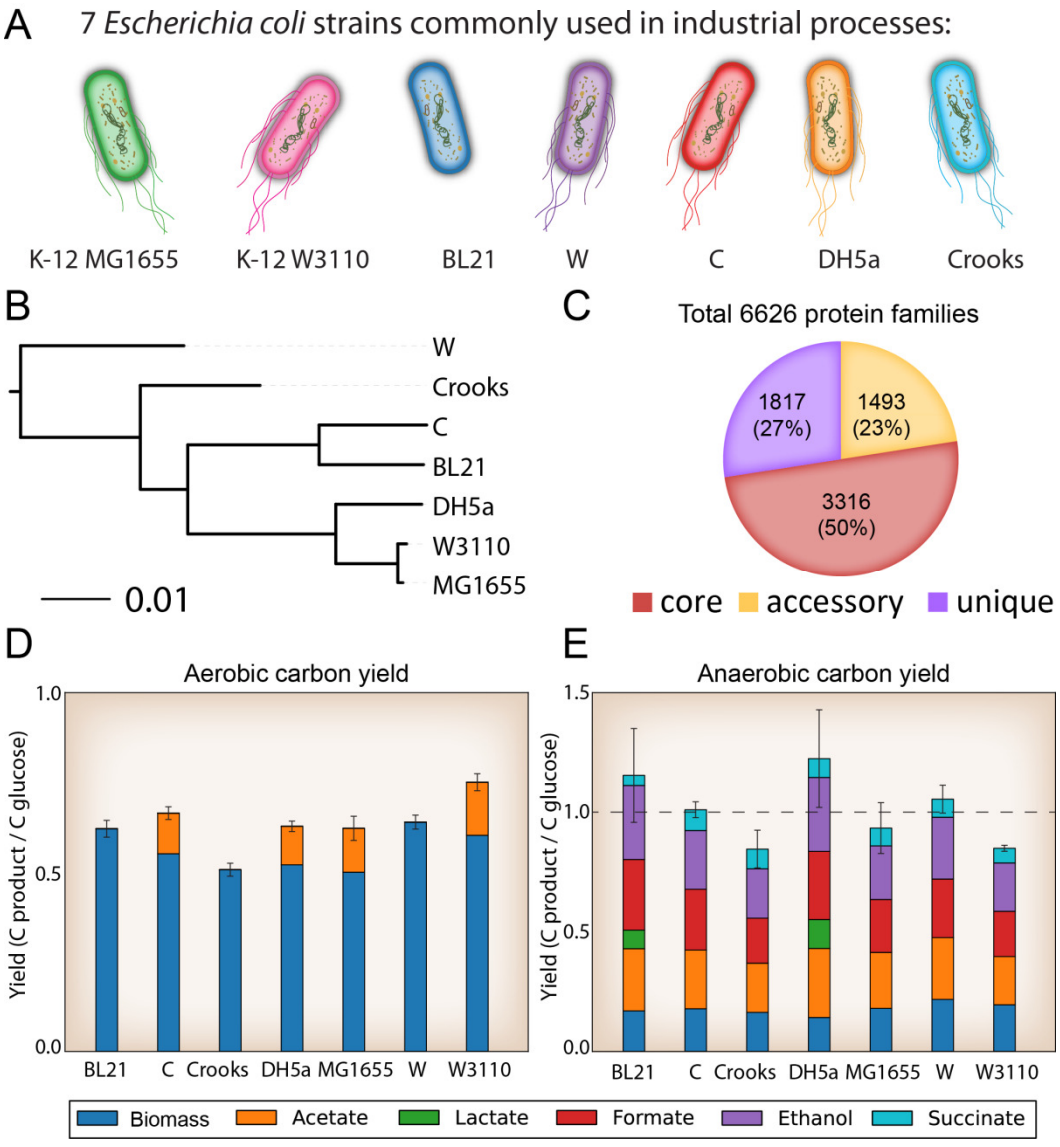
production. R-scores for each strain's amino acid production potential were averaged. Amino acid production R-scores greater than 0.5 standard deviations above the mean were considered highly produced. A sensitivity analysis on this cutoff was performed with cutoffs ranging from 0.25-1$\sigma>\mu$ examined. These different cutoffs did not change the overall results of this analysis. The overlap between demanded amino acids for each construct and amino acids highly produced by each strain were determined to predict which strain might be best suited for production of a given construct. In aerobic conditions it was found that DH5a was often the best producer while in anaerobic conditions Crooks was often best (see main text).

5.6 Acknowledgements

The work was funded by the Novo Nordisk Foundation. JM was funded by grant 1R01GM057089 from the NIH/NIGMS.

Chapter 5, in part, is a reprint of the material Monk, JM, Koza A, Campodonico A, Machado D, Seoane JM, Palsson BO, Herrgård MJ, Feist AM. Quantifying variation within the bacterial species *E. coli* and the impact on host strain selection. *In Preparation.* The disseration author was the primary author of this paper.

Figure 5.1. Statistics of the 7 strains
A) The seven industrially-relevant *E. coli* strains selected for this study. This panel establishes the colour scheme that will be used to represent strain-specific data on these strains throughout the rest of the manuscript, as well as the motility characteristics of each strain. **B)** A phylogenetic tree based on full-genome DNA alignment of the seven strains. **C)** Total protein families in the strains examined. Core genes (red) are those that were present in all 7 strains, genes present in 2-6 strains are labelled accessory (yellow), and unique genes are only present in a single strain (purple). **D & E)** Physiological behaviour and comparison of the different strain's carbon yield in aerobic and anaerobic growth conditions. Yields are calculated in terms of glucose uptake rates. Carbon dioxide was not measured. Overall, the by-product profiles differed across the strains and some anaerobic yields are not fully captured in the by-products measured. This is likely due to CO_2 evolution from formate dehydrogenase.


A
7 Escherichia coli strains commonly used in industrial processes:
K-12 MG1655
K-12 W3110
BL21
W
C
DH5a
Crooks
B
W
Crooks
C
BL21
DH5a
W3110
MG1655
0.01
C
Total 6626 protein families
1817 (27%)
1493 (23%)
3316 (50%)
core
accessory
unique
D
Aerobic carbon yield
Yield (C product / C glucose)
1.0
0.5
0.0
BL21
C
Crooks
DH5a
MG1655
W
W3110
E
Anaerobic carbon yield
Yield (C product / C glucose)
1.5
1.0
0.5
0.0
BL21
C
Crooks
DH5a
MG1655
W
W3110
Biomass
Acetate
Lactate
Formate
Ethanol
Succinate

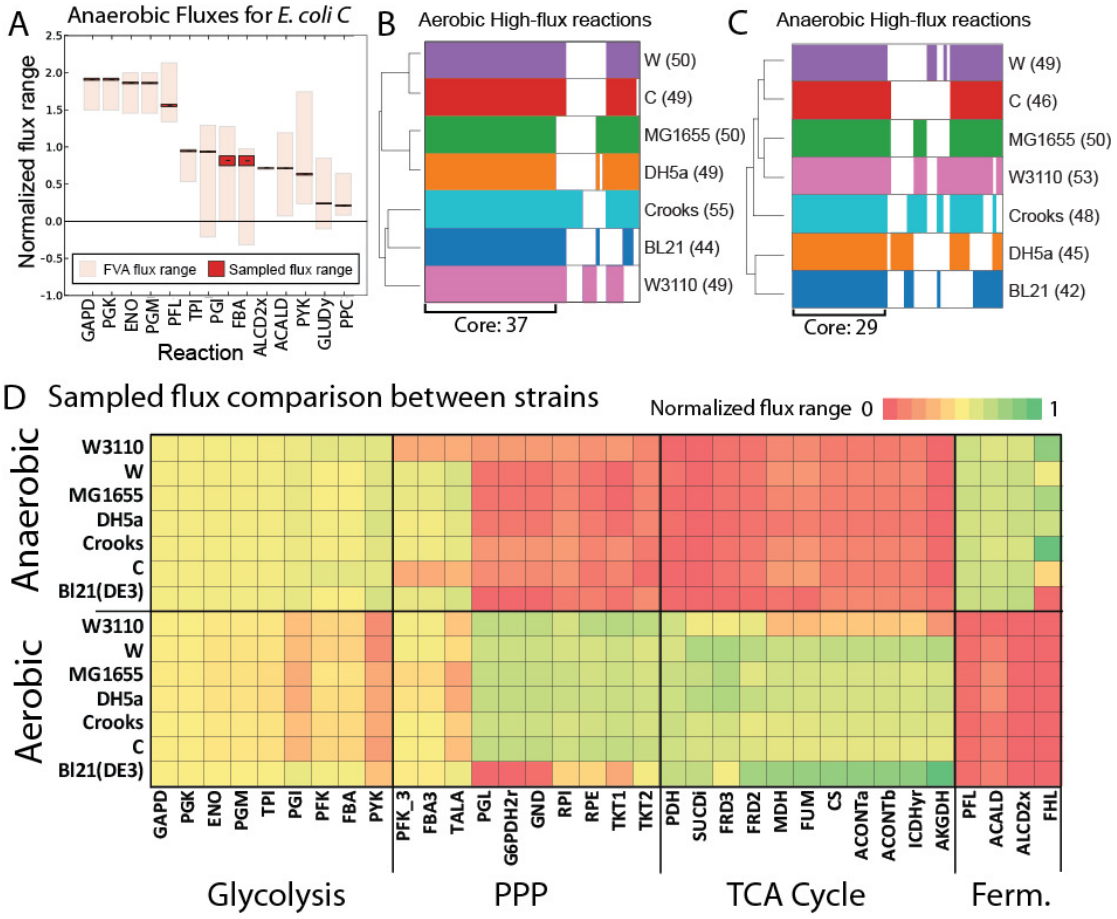


Figure 5.2. Computationally determined high flux reactions from physiological data.
A) Each strain-specific model was constrained using measured physiological data. Flux variability analysis and sampling were performed on each constrained model. All fluxes were normalized to glucose uptake rate and sorted by mean sampled flux value. All graphs for each strain and each growth condition are available in **Supplementary Figure 7**. The absolute values of normalized flux values were log transformed. High flux reactions were determined to be those reactions with sampled-flux values greater than 1.5 standard deviations above the mean of all sampled flux values. **B and C)** High flux reactions for each strain were clustered and plotted. The counts of high flux reactions for each strain are indicated next to the strain name in parentheses. Shared (core) and unique high flux reactions for each strain are shown for aerobic and anaerobic conditions. **D)** Sampled flux values were compared between the strains in aerobic and anaerobic conditions to highlight condition- and strain-specific behaviour. Reaction abbreviations are given in **Supplementary Data File 3**.

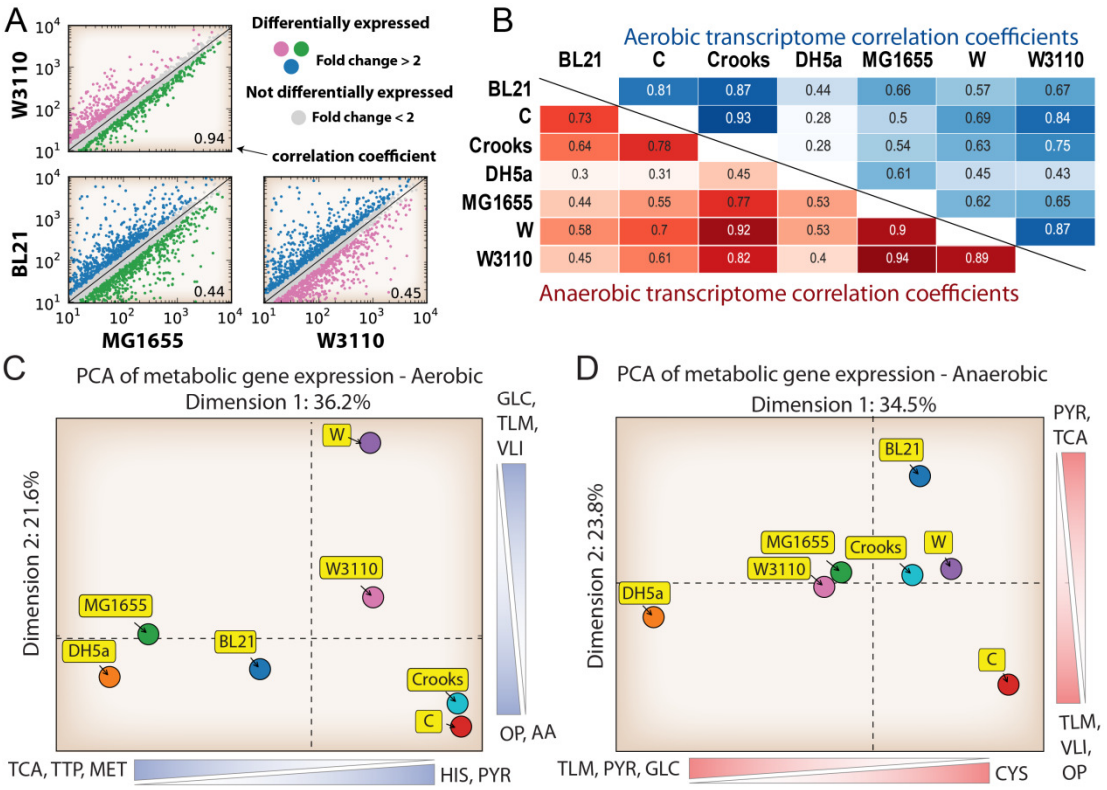


Figure 5.3. Gene expression analysis.
A) Example of a comparison of the transcripts levels between strains W3110, MG1655, and BL21, as well as their correlation coefficient. Strains MG1655 and W3110 have a 0.94 correlation coefficient for expression of shared genes. In contrast, strain BL21 shows divergent gene expression compared to strains MG1655 and W3110 with a much lower correlation coefficient of 0.44 and 0.45, respectively. **B)** The pairwise correlation coefficients for each strain in both aerobic (top, green) and anaerobic (bottom, red) conditions. **C)** PCA plot of expression values for shared metabolic gene expression between strains aerobically and **D)** anaerobically. Scale bars represent metabolic subsystems that majorly contribute to the given dimension. Abbreviations: TCA: Citric Acid Cycle, APM: Arginine and Proline Metabolism, GLC: Glycolysis/Gluconeogenesis, MET: Methionine Metabolism, HIS: Histidine Metabolism, OP: Oxidative Phosphorylation, TLM: Threonine and Lysine Metabolism, VLI: Valine, Leucine and Isoleucine Metabolism, PYR: Pyruvate Metabolism, TTP: Tryptophan, Tyrosine and Phenylalanine Metabolism, AA: Aspartate and Alanine Metabolism.

Figure 5.4. A comparison of high flux reactions and highly expressed metabolic genes.
A) The occurrence of a given gene and the reaction it catalyses falling in the intersection of both high flux and highly expressed sets for the seven strains examined. Several pathways are enriched in this intersection set (e.g., lower glycolysis). Genes were grouped if they had identical counts under aerobic and anaerobic conditions. There were several genes that were in this intersection exclusively for aerobic or anaerobic conditions (right side of graph). **B & C)** The measured physiological data was integrated with genome-scale models to predict changes in gene expression during a shift from aerobic to anaerobic conditions. Shown are two examples for **(B)** malate dehydrogenase (MDH) and **(C)** the Electron Transport System. A map is shown of the predicted reaction and its neighbours (left) along with the prediction of the intracellular flux change between aerobic and anaerobic conditions (middle). The actual measured fold change in expression is graphed (right) for comparison with model-predicted flux changes (dashed lines indicate a threshold for minimum magnitude and * indicates significance for the predicted and measured changes). The example for prediction of *mdh* **(B)** demonstrates that the model correctly predicts a change in expression; all 7 models predicted a change in MDH flux that exceeded the minimum threshold and the measured expression change of *mdh* in each strain was large and significant. The *nuo* genes **(C)** that catalyse the NADH dehydrogenase reaction also demonstrate good model-based prediction of expression change.

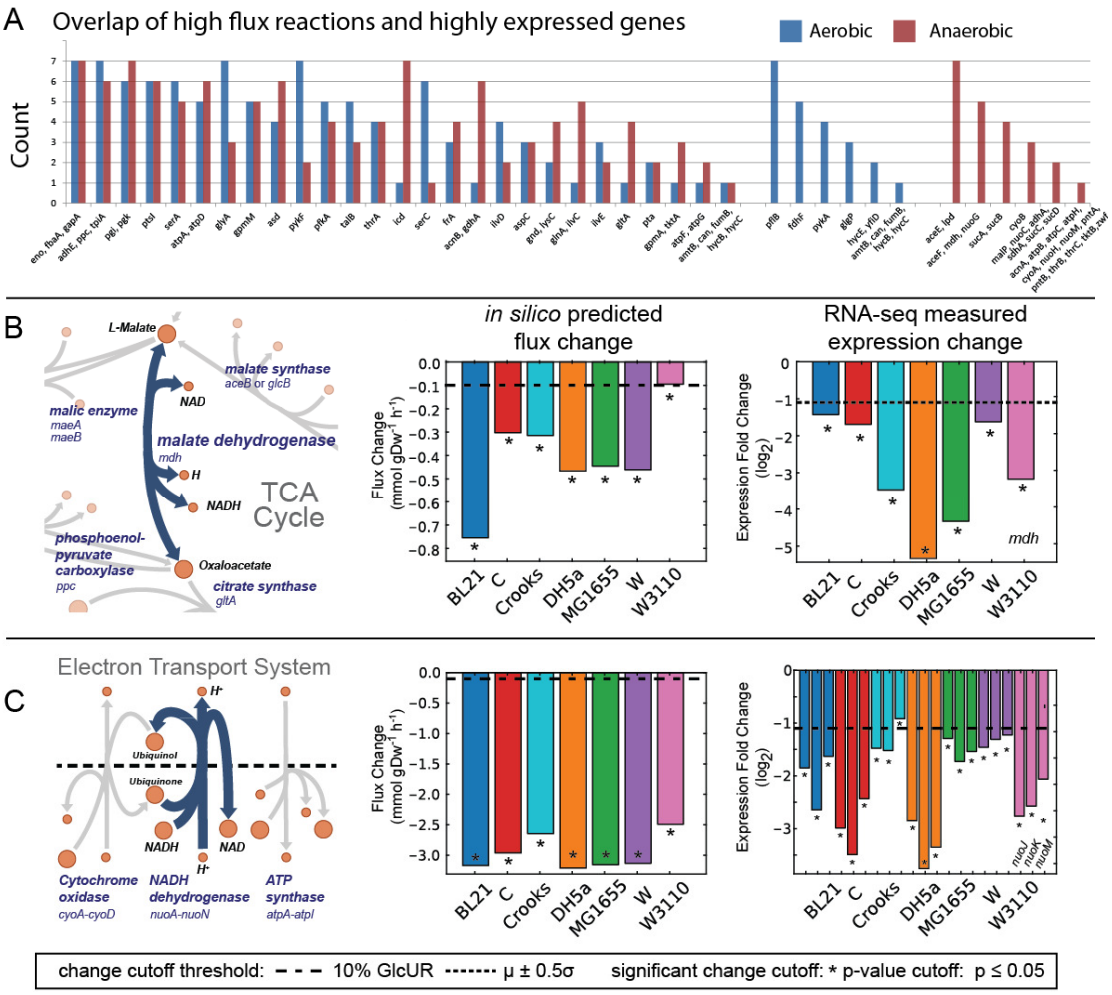
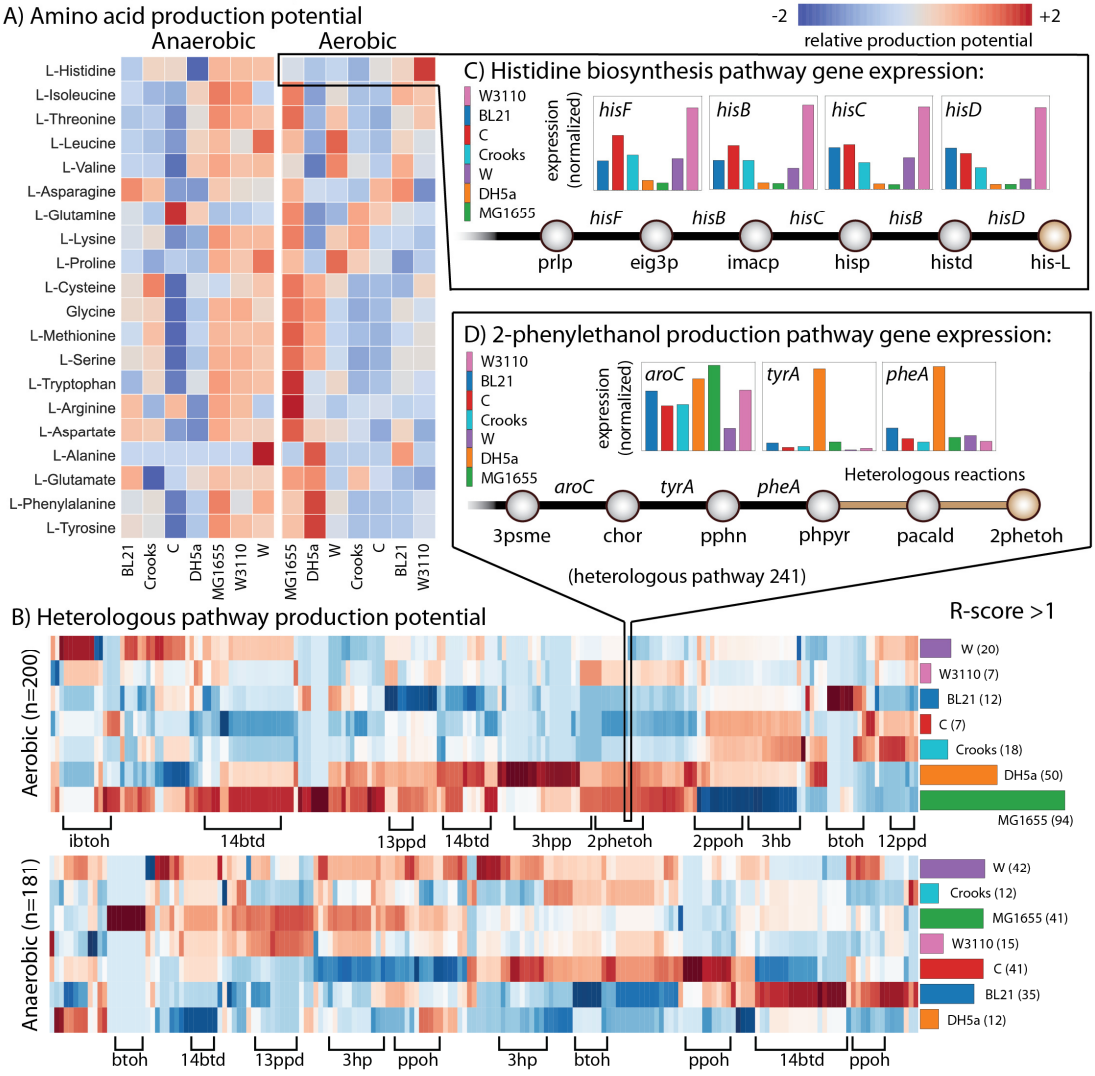

A
Overlap of high flux reactions and highly expressed genes
Aerobic
Anaerobic
Count
B
L-Malate
malate synthase
aceB or glcB
NAD
malic enzyme
maeA
maeB
malate dehydrogenase
mdh
H
NADH
TCA Cycle
phosphoenol-pyruvate carboxylase
ppc
Oxaloacetate
citrate synthase
gltA
in silico predicted flux change
RNA-seq measured expression change
Flux Change (mmol gDw-1 h-1)
Expression Fold Change (log2)
mdh
BL21
C
Crooks
DH5a
MG1655
W
W3110
C
Electron Transport System
H+
Ubiquinol
Ubiquinone
NADH
NAD
Cytochrome oxidase
cyoA-cyoD
NADH dehydrogenase
nuoA-nuoN
ATP synthase
atpA-atpI
nuoJ
nuoK
nuoM
change cutoff threshold: - - - 10% GlcUR ------- μ ± 0.5σ significant change cutoff: * p-value cutoff: p ≤ 0.05

Figure 5.5. Strain-specific production potential.
The product potential for targeted metabolites was evaluated based on native gene expression for high yield pathways of interest measured using RNAseq and genome-scale modelling (see main text and methods). **A)** A heat map displaying the relative production potentials for all 20 amino acids (left axis) for each strain (bottom axis) in aerobic and anaerobic conditions. Red indicates the highest potential and blue lowest (see legend). **B)** A heat map of the relative production potential for 245 and 200 viable heterologous pathways in aerobic (top) and anaerobic (bottom) conditions, respectively. Heterologous pathways are clustered (columns) based on the target product (there can be many compounds for a given compound) and some of the most abundant are labelled on the bottom axis. The right axis shows a plot of the number of instances where each strain has an 'R-score' (relative production potential score, see methods) > 1. **C)** Example demonstrating the production potential for histidine biosynthesis. Shown are the final five reaction steps and relative expression levels of their catalysing genes for each strain. Strain W3110 (pink) has greater gene expression of these *his* operon genes, making it particularly well suited to produce histidine. **D)** A similar example demonstrating production potential for heterologous production of 2-phenylethanol. Here, the last three steps (before the heterologous pathway) are shown with their relative native expression levels. This heterologous pathway branches from phenylpyruvate (phpyr), an intermediate of tyrosine biosynthesis. DH5a has high native expression of these 3 steps along with others in the pathway. All metabolite abbreviations are listed in **Supplementary Data File 3**. The scores for each product and contributing expression values and flux profiles are available in **Supplementary Data File 15**.


A) Amino acid production potential
Anaerobic
Aerobic
-2
+2
relative production potential
L-Histidine
L-Isoleucine
L-Threonine
L-Leucine
L-Valine
L-Asparagine
L-Glutamine
L-Lysine
L-Proline
L-Cysteine
Glycine
L-Methionine
L-Serine
L-Tryptophan
L-Arginine
L-Aspartate
L-Alanine
L-Glutamate
L-Phenylalanine
L-Tyrosine
BL21
Crooks
C
DH5a
MG1655
W3110
W
C) Histidine biosynthesis pathway gene expression:
W3110
BL21
C
Crooks
W
DH5a
MG1655
expression (normalized)
hisF
hisB
hisC
hisD
prlp
eig3p
imacp
hisp
histd
his-L
D) 2-phenylethanol production pathway gene expression:
aroC
tyrA
pheA
Heterologous reactions
3psme
chor
pphn
phpyr
pacald
2phetoh
(heterologous pathway 241)
B) Heterologous pathway production potential
R-score >1
Aerobic (n=200)
W (20)
W3110 (7)
BL21 (12)
C (7)
Crooks (18)
DH5a (50)
MG1655 (94)
ibtoh
14btd
13ppd
14btd
3hpp
2phetoh
2ppoh
3hb
btoh
12ppd
Anaerobic (n=181)
W (42)
Crooks (12)
MG1655 (41)
W3110 (15)
C (41)
BL21 (35)
DH5a (12)
btoh
14btd
13ppd
3hp
ppoh
3hp
btoh
ppoh
14btd
ppoh

Chapter 6 Optimizing genome-scale network reconstructions

6.1 Abstract

Genome-scale Network Reconstructions (GENREs) are organism-specific knowledge-bases. GENREs can be converted to mathematical models that enable computation of an organism's metabolic capabilities and phenotypic expression. In conjunction with complex omics datasets, GENREs can be applied to solve basic and applied biological questions. Here, we present an analysis of the metabolic reaction lists (so called reactomes) and phylogenetic coverage of current GENREs, showing that their scope and content is more limited than generally appreciated. The causes and consequences of their limited scope are discussed and strategies for the continued development of the field are suggested.

6.2 Introduction

A GENRE is built systematically through a quality controlled bottom-up workflow by using genome annotation, omics data sets, and legacy knowledge (Orth et al., 2010c). Thus, GENREs should embody the best representation of the metabolic capabilities of a target organism based on the information available at the time of reconstruction. GENREs are knowledge-bases that integrate and reconcile genome annotation, legacy biochemical data, computational simulations, and phenotypic expression. They allow researchers to collaborate, test, and readily share new hypotheses about metabolic functions in a target organism. As a result, interest in network reconstructions as well as the scope of their applications has grown continually (Österlund et al., 2012, Bordbar and Palsson, 2012, Kim et al., 2012, Lewis et al., 2012b, McCloskey et al., 2013d).

The ability to perform genome-scale network reconstruction has progressed rapidly. The first GENRE was constructed for *Haemophilus influenza* in 1999(Edwards

and Palsson, 1999) just a few years after the first whole genome sequence appeared in 1995(Fleischmann et al., 1995). This initial reconstruction represented the development of a conceptual basis for forming and building GENREs and demonstrated that the metabolic genotype-to-phenotype relationship could be computed from a mechanistic and a genome-scale basis. Following this initial reconstruction, best practices for generating metabolic reconstructions were developed based on experience with highly studied and well-characterized model organisms. Iterative updates to the GENRE for *Escherichia coli* created standards for the gene-protein-reaction association, elemental and charge balancing (Edwards and Palsson, 2000a) and addition of thermodynamic information (Feist et al., 2007). Updates to the GENRE for *Saccharomyces cerevisiae* standardized cellular compartmentalization (Forster et al., 2003). Thus, high-quality protocols for metabolic reconstruction were established[1], and the development of GENREs progressed to the successful reconstruction of human metabolism (Duarte et al., 2007b), photosynthesis (Nogales et al., 2012), and light-driven metabolism (Chang et al., 2011). These developments have led to a multitude of successful applications reviewed elsewhere (Österlund et al., 2012, Bordbar and Palsson, 2012, Kim et al., 2012, Lewis et al., 2012b, McCloskey et al., 2013d).

Over the last five years the number of new GENREs has grown rapidly (**Fig. 1a**) and expanded the 'metabolic space' suitable for computational analysis (Kim et al., 2012). Further, automated reconstruction approaches are now available to create draft reconstructions, reducing the time and effort required to make a metabolic reconstruction(Henry et al., 2010b, Vitkin and Shlomi, 2012, Agren et al., 2013b). GENREs have become accepted as valuable tools to teach and analyze biological processes at the systems level (Rabinowitz and Vastag, 2012). Therefore, more than a

decade after the publication of the first GENRE, it is timely to analyze the metabolic knowledge represented in published network reconstructions to assess the overall progress and status of this field.

6.3 Coverage of metabolic reactions

While the metabolic network reconstruction field may appear mature, there are still many challenges that need to be addressed. The metabolic scope and coverage of GENREs has not progressed in accordance with their rising number. An analysis of the number of new metabolic reactions that have been incorporated into new GENREs in recent years shows that just a few reconstructions dominate the reaction list while most of the GENREs published have not contributed significantly to the represented metabolic space **(Fig 1)**. This limited representation is further demonstrated by comparing the BRENDA (Schomburg et al., 2013) enzyme database to enzymatic activities found in current GENREs. Surprisingly, just 33% of the enzymatic activities in BRENDA assigned to metabolism are included in the group of GENREs that we analyzed. While this result could be biased due to incomplete mapping or redundant EC nomenclature, the small portion of the metabolic knowledge currently included in existing GENREs indicates incomplete coverage of known metabolic reactions.

Since many new GENREs are based on existing ones, questions arise regarding their independence and contribution to new knowledge. If the metabolic knowledge included in a GENRE indeed reflects the metabolic capabilities of the target organism, we would expect clustering of content that mimics the evolutionary trajectory of different organisms. However, a similarity analysis of GENRE reaction content shows that's not the case. Multiple Correspondence Analysis (MCA) of the content of 53 (**Supplementary Table 2 and File 1**) curated GENREs out of the 117 published to date

(http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms) (**Supplementary Table 1**), shows that a surprisingly high degree of similarity is found among most of existing GENREs **(Fig. 1b)** regardless of their location in the phylogenetic tree (**Fig. 2**). Many GENREs cluster close to the center of the diagram showing that reconstructed organisms as metabolically diverse as *Pseudomonas aeruginosa*, *Staphylococcus aureus, Clostridium beijerinckii, and Synechocystis sp.* PCC 6803 have similar reaction content (**Fig. 2b**). This indicates that the metabolic space of currently published GENREs is limited to well-conserved metabolic pathways rather than being a comprehensive representation of the biochemical capabilities of these target organisms. This bias leads to over-representation of primary metabolic pathways in GENREs relative to the important aspect of secondary metabolism, which often defines the signature identity of a target organism.

In contrast, three groups of GENREs, dominated by enterobacteria, yeasts and photosynthetic eukaryotes were distinguishable in their reactomes. Whereas the first GENRE of *E. coli, i*JE660(Edwards and Palsson, 2000a), is in the cluster at the center of the diagram, further updates *i*JR904(Reed et al., 2003), *i*AF1260(Feist et al., 2007), and *i*JO1366(Orth et al., 2011) have moved steadily away from the center based on increasing organism-specific metabolic content **(Fig 2c)**. A similar path can be observed for *S. cerevisiae* from the initial GENRE, iFF708 (Forster et al., 2003), to the latest version YEAST5 (Heavner et al., 2012). Given these two examples, it seems reasonable to expect that iterative reconstruction of a target organism should bring out its unique characteristics and distinguish them from the rest. Detailing species-specific reactomes may thus be a way to define biological diversity.

Beyond reconstruction content, the range of organisms for which GENREs exist is limited, raising questions about the breadth of coverage of metabolism across the biosphere. The NCBI taxonomy database can be used to examine the phylogenetic distribution of published GENREs (**Supplementary Table 1**). This examination reveals that while some phyla are well represented by multiple GENREs, there are many others without any reconstructions (**Fig. 3**). For example, there are 32 metabolic reconstructions of species that are members of the proteobacteria phylum. This represents over 40% of the 77 reconstructed species to date. In stark contrast, there are 15 phyla containing species with sequenced genomes that have no reconstruction. These unrepresented phyla fall across the tree of life in all kingdoms; bacteria, archaea, and eukaryota. Thus, the narrow phylogenetic coverage of currently available GENREs is an insufficient representation of the metabolic capabilities found on earth and it is evident that further reconstructions of diverse organisms across the tree of life are needed in order to achieve broad and comprehensive phylogenetic coverage.

In summary, it may be concluded that new organisms across the tree of life need to be reconstructed and that many current metabolic network reconstructions are still lacking a significant portion of their target organism's reactome, suggesting that they are still in a development stage similar to the first reconstruction of *E. coli*, *i*JE660 published thirteen years ago.

6.4 Multiple limitations hamper the full development of the genome-scale network reconstruction field

The lack of comprehensive biological knowledge is primarily responsible for the limited metabolic coverage in current GENREs (**Fig. 4**). Even for a microorganism as well studied as *E. coli,* only half (~54%) of the protein-coding gene products have direct

experimental evidence for their function(Riley et al., 2006), and up to one-third of its proteome remains functionally un-annotated(Hu et al., 2009). This limitation has come into focus in recent years when extending network reconstruction to less-studied organisms with a low species knowledge index (SKI)(Janssen et al., 2005). The SKI is defined as the ratio of publications divided by the number of protein-encoding genes for a given organism. GENREs for organisms with low SKI are often poorly curated and validated due a distinct lack of biological knowledge. Analogously to mistakes made with automated genome annotation(Schnoes et al., 2009), the inclusion of an incorrect gene-to-protein-to-reaction (GPR) and/or wrong reaction in a GENRE can be disseminated to a new reconstruction.

The lack of biological knowledge about a target organism is not the only limitation leading to poor coverage of existing metabolic reconstructions. Surprisingly, the amount of biochemical information available for a given organism seems to have a relatively small impact on the final metabolic coverage of its GENRE. Species with high SKI values can still have GENREs with limited metabolic coverage (**Fig. 2**). This limitation may indicate under-utilization of existing biological knowledge when performing the reconstruction. The primary reason for such under-utilization is a lack of substantial manual curation, most likely due to the extensive manual labor and resources required. Thus, genome annotation is often the main source of content in new GENREs. However, genome annotation cannot be considered a complete and accurate source of biochemical information(Schnoes et al., 2009). In addition there are often a large number of un-annotated protein-coding genes. This excludes a large portion of the metabolic space because 30-40% of known enzymatic activities are global orphans (i.e., no genes encoding their activity have been discovered)(Lespinet and Labedan, 2005,

Chen and Vitkup, 2007, Pouliot and Karp, 2007). Thus, the dominance of genome annotation over others sources of information and the disregard of legacy data on biochemical and/or physiological studies, leads to a significant underrepresentation of the metabolic potential of the target organism.

The lack of inclusion of legacy data in the reconstruction process is particularly apparent for secondary metabolism. While central metabolism may be largely conserved, secondary metabolism tends to be organism-specific and thus much more difficult to reconstruct based on genome annotation alone. As a result, secondary metabolism is often passed over during the reconstruction process. These unrepresented pathways often include the biosynthesis of cofactors and vitamins, lipids, cell envelope and organism-specific pathways**.** This impacts the Biomass Objective Function (BOF) that is used to define cellular growth requirements(Feist and Palsson, 2010c). Unfortunately, incomplete biosynthetic demands lead to simplified BOFs unrepresentative of the organism's true physiology. This widespread practice reduces the computable metabolic space by creating blocked reactions(Orth and Palsson, 2012). The consequence is an inability to use GENREs for fundamental systems biology studies like gene essentiality prediction and omics data contextualization for the target organism. Often, the lack of biological knowledge and insufficient legacy data curation are partially mitigated through the use of high-throughput technologies. However, due to the high amount of resources required, the application of high-throughput datasets for modeling is not readily accessible to individual reconstruction efforts, and these technologies remains untapped in metabolic reconstructions.

The lack of rigor in applying well-established reconstruction protocols is another contributing factor to the limited metabolic coverage of current GENREs. Despite the fact

that standards for reconstruction exist, GENREs that include unbalanced mass and charge reactions, lumped reactions instead of complete pathways, and incompletely compartmentalized networks are still being published. For example, it is striking that from 54 reconstructions of Gram negative bacteria, just 11 (or 20%) include periplasm as a cellular compartment.

Another weakness of the reconstruction process beyond biological limitations is the lack of a standardized representation for common metabolites and reactions included in current GENREs. This shortcoming makes metabolic reconstructions unintelligible and impedes automated omics data mapping. It also negates the benefits that could be achieved by comparative analysis of GENREs, which is the main reason only 53 published reconstructions were compared in this analysis. Clearly, the field needs to recognize these limitations, and continue to develop and adhere to best practices. The publication of poor quality reconstruction does not advance the field as a whole, and does injustice to a systems biology study of the metabolism of the target organism.

6.5 Towards comprehensive metabolic coverage and broader deployment of GENREs

Thus we need to consider how to improve the status of the field, to advance it and broaden its scope. We discuss three critical issues towards this end. *1. Targeted application of high-throughput technologies.* To mature the metabolic reconstruction field we can increase biological knowledge by carefully applying high-throughput technologies. Several recent studies have used targeted high-throughput metabolomic, transciptomic and mutant screen data sets as well as computational structure and

metabolite docking predictions to discover new functions(Li et al., 2010, Baran et al., 2013)' (Nakahigashi et al., 2009a).

There is a need to develop high-throughput reactome determination technologies, however unfortunately large-scale biochemistry presents a great challenge. Efforts aimed at determining metabolite-protein interactions have shown promise in addressing this challenge. A systematic large-scale investigation of *in-vivo* protein-metabolite interactions in yeast has been developed(Li et al., 2010) leading to the discovery of several new metabolite-protein regulatory interactions. This technology could also apply to discovering metabolic interactions such as cofactors with enzymes and energy sources used in novel biochemical reactions. Another approach, untargeted metabolomics, has been used to assign function to new genes in a high-throughput manner (Baran et al., 2013).

The reconstruction process itself is an exciting opportunity to increase the biological knowledge for the target organism. Because GENREs represent a biochemically and genetically structured knowledge-base, they can be queried and interrogated using *in-silico* analytical methods. Reconstruction is an iterative process where errors in model prediction drive new hypotheses(Orth and Palsson, 2012) (**Fig. 4**). A powerful example of using a GENRE to interpret experimental results and then discover novel biochemistry has been illustrated(Nakahigashi et al., 2009a). This study combined GENREs with systematic multiple gene knockout strains to discover new reactions carried out by phosphofructokinase and aldolase, two extensively studied enzymes in *E. coli* glycolysis. Experimental results were compared to computational predictions and disagreements suggested missing reactions in the reconstruction. The putative reactions were then confirmed using metabolome analyses and *in vitro*

enzymatic assays illustrating that even for extensively studied areas of metabolic biochemistry, there is still more to learn. Combining GENREs with the high-throughput techniques above will further improve discovery of new biological functions in a synergistic manner. *2. Building high quality reconstructions with community participation and buy-in*. As noted above, building a high quality reconstruction is non-trivial given that a high-quality reconstruction relies on extensive manual curation of legacy datasets, minute attention to biochemical and molecular detail and careful phenotyping. Accurate and complete GENRE development is a multidisciplinary activity. It requires community buy-in from domain experts in diverse disciplines. An ideal team would combine strong biological knowledge of the target organism with access to legacy data. Following these guidelines, reconstruction jamborees have been carried out with success for three target organisms (*Saccharomyces cerevisiae* (Herrgard et al., 2008), *Salmonella typhimurium* LT2 (Thiele et al., 2011), and *Homo sapiens* (Thiele et al., 2013a)). Further and more structured efforts are needed to curate existing content and to expand the scope of GENREs, potentially through a "crowd-sourcing" mechanism where multiple individuals can contribute to a reconstruction so that it contains as much legacy data as possible.

In order to form such teams, the reconstruction community must reach out to domain experts, many of whom are currently unfamiliar with the metabolic reconstruction process. Recently, a multidisciplinary team of researchers (that included experts from Pharmaceutical Chemistry, Genomic Biology, Biochemistry, Bioengineering, Chemistry and Microbiology departments) used protein structure and genome context to functionally annotate new enzymes in *Pelagibaca bermudensis(Zhao et al., 2013)*. Computational analysis of metabolite docking to three-dimensional structures

(experimentally derived or homology-based) was used to predict substrate-specificities of several enzymes in a new catabolic pathway. Recent studies have integrated protein structure information into GENREs(Chang et al., 2013a), thus opening GENREs to similar analyses. This success should strongly encourage similar efforts for metabolic reconstruction. Such multi-disciplinary teams should be motivated by the broadening appreciation of the power of GENREs and the likelihood of increased prestige of their publication venue.

3. Increase the coverage of phylogenetic tree. Above we observed that metabolic reconstructions do not exist across the tree of life. Ignoring a major portion of living organism is a limitation that prevents full maturity of the field. If we wish to understand and study the metabolic capabilities resident on earth, reconstruction efforts must be undertaken for diverse organisms spread throughout the tree of life analogous to the GEBA project to sequence genomes of diverse organisms (Wu et al., 2009). Those organisms down branches of the tree of life where no reconstructions currently exist, but for which biochemical and legacy data exists, should be targeted first. Once high-quality reconstructions are completed for such target organisms, the content can be mapped to closely related species, akin to what has been done to generate reconstructions for *Klebsiella (Liao et al., 2011a)*, *Yersinia (Charusanti et al., 2011)* and *Salmonella (Thiele et al., 2011)* based on the *E. coli* reconstruction(Feist et al., 2007).

6.6 Outlook

Genome-scale metabolic reconstructions and modeling represents an advance in genome-scale science and systems biology. They allow for the study of living organisms as systems through the integration and contextualization of myriad high-throughput

experimental and computational data sets. GENREs allow us to globally examine our understanding and knowledge of metabolism for a target organism.

GENREs in their current incarnation have enabled a surprisingly wide range of basic and applied biological studies(Österlund et al., 2012, Bordbar and Palsson, 2012, Kim et al., 2012, Lewis et al., 2012b, McCloskey et al., 2013d). Although notable successes have been achieved, the field is still immature. Most current metabolic reconstructions cannot strictly be considered genome-scale, but instead models of primary metabolism that may be unsuitable for deeper systems biological studies of the target organism. We need to undertake a concerted effort to improve metabolic coverage of well-studied organisms and to capture known metabolic capabilities in the various branches of the phylogenetic tree. Furthermore, as anticipated over 10 years ago(Reed and Palsson, 2003b), GENREs can be expanded to include other cellular processes such as transcription and translation(Thiele et al., 2009a, Lerman et al., 2012b), transcriptional regulation(Covert et al., 2004), and metabolic maintenance functions(Linster et al., 2013). More comprehensive inclusion of such processes and their seamless integration with metabolism would allow for an assessment of their quantitative interrelations. But high-quality metabolic reconstructions must be a prerequisite.

GENREs are foundational to the formulation of quantitative genotype-phenotype relationships and thus the more comprehensive and high quality a GENRE is the more phenotypic functions can be computed from the corresponding genome-scale model. Increased scope and quality of computed phenotypic functions and their experimental validation in turn steadily increases our understanding of the functions of the target organism, and thus the underlying multi-scale relationship between the genotype and the

phenotype. Such understanding will be a key to solving many basic and applied biological challenges that lie ahead.

6.7 Acknowledgements

We would like to thank Benjamin Heavner, Aarash Bordbar, Adam Feist and Joshua Lerman for helpful discussions and critical review of the manuscript. JM is funded by NIH R01 GM057089. JN was funded by the Spanish Ministry of Education through the National Plan for Scientific Research, Development, and Technological Innovation 2008–2011.

Chapter 6, in part, is a reprint of the material Monk JM*, Nogales J*, Palsson BO: Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014, 32(5):447-452. The disseration author was the primary author (equally contributing with Juan Nogales) of this paper.

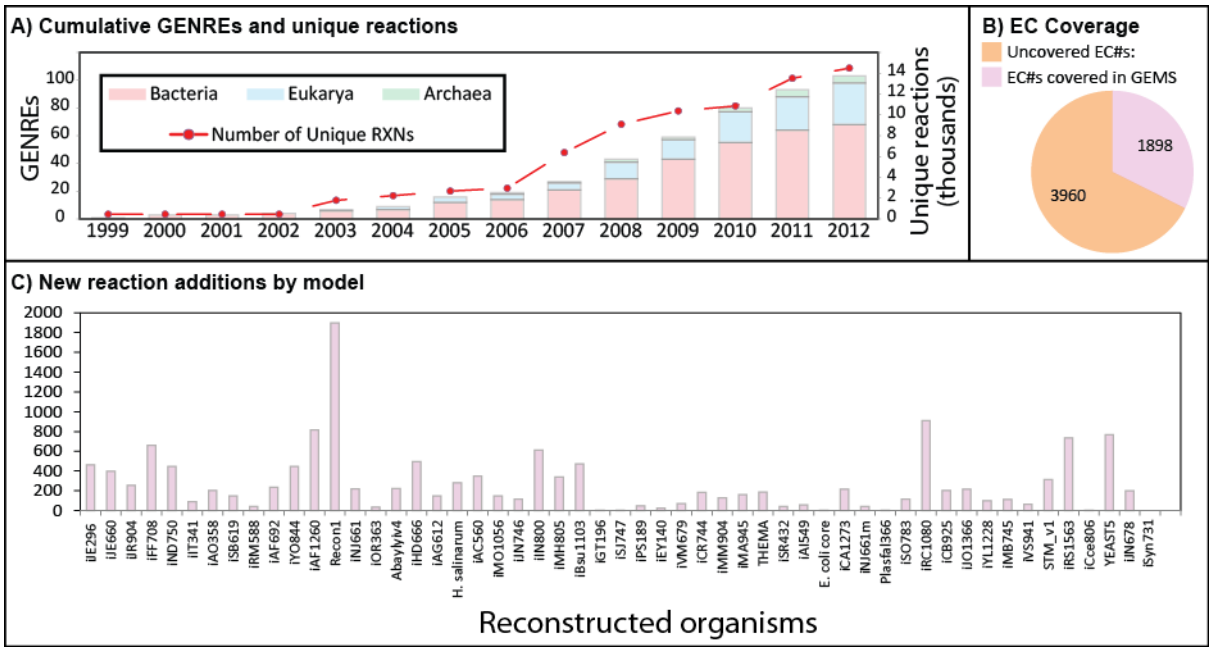


Figure 6.1. Evolution of metabolic networks and global reactome coverage over time. **(A)** The cumulative number of GENREs (bar chart) and unique reactions (line chart) belonging to the current metabolically computable space. **(B)** Current coverage of EC (Enzyme Commission) numbers in published GENREs. **(C)** Contribution to the coverage of metabolic space per GENRE, as determined by the number of unique reactions added at the time of publication.

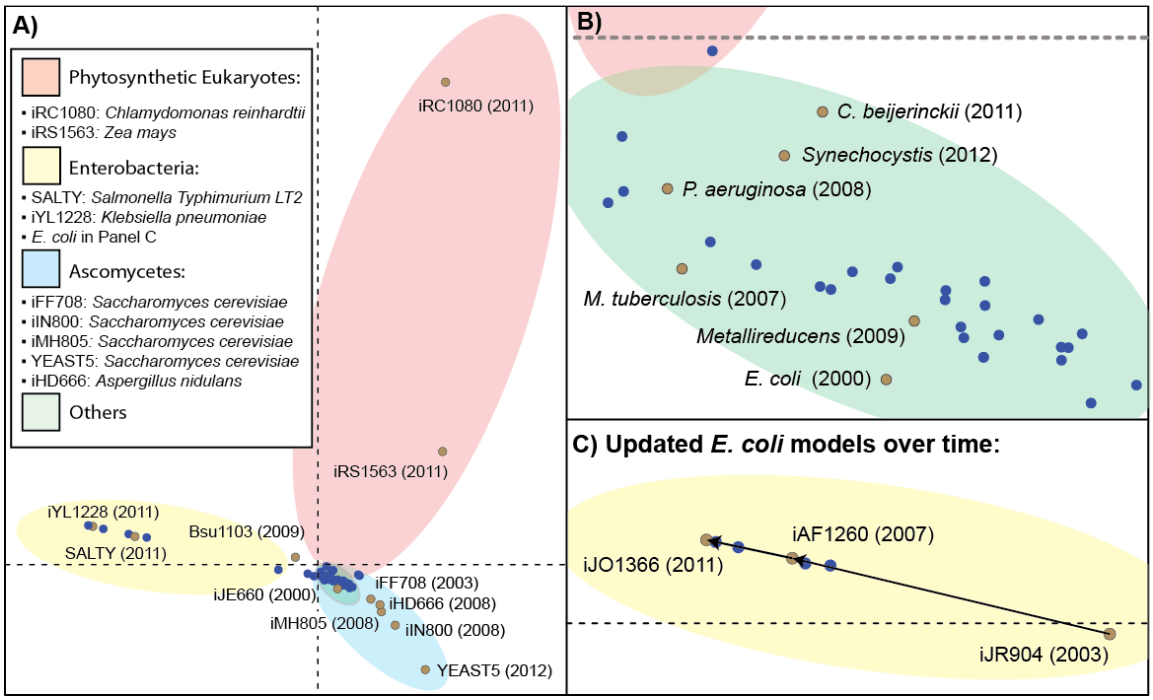


Figure 6.2. Multiple Correspondence Analysis showing similarity between current GENREs.

Of 117 published curated GENREs (as of February 2013) up to 53 could be consistently represented and subsequently analyzed. The reason for the incomplete analysis was the unavailability of several metabolic reconstructions as well as the high disparity of nomenclatures for reactions and metabolites(Ganter et al., 2013). The human GENRE (Recon 1) was removed from this analysis because it is significantly different from the rest of the GENREs analyzed. A blue dot represents each GENRE. The tan dot indicates selected labeled models (**see SBRG website(http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms) and SI**). The axes represent degree of similarity, for example GENREs in the blue ellipse are highly similar to each other, but very different from those in the yellow ellipse. **A)** The GENREs clustered in four different groups. Most grouped close to the center of coordinates reflecting minimal differences between them (green). However, enterobacteria (yellow), yeasts (blue), and photosynthetic eukaryotes (pink) grouped far away, indicating that these reconstructions are very different from each other and have content that covers a different section of the metabolic space. **(B)** Detail of the main group of GENREs. The tan circles represent metabolically different organisms with GENREs that clustered together. **C)** Detail of the trajectory over time of multiple iterations of GENREs for the model organism *E. coli*.

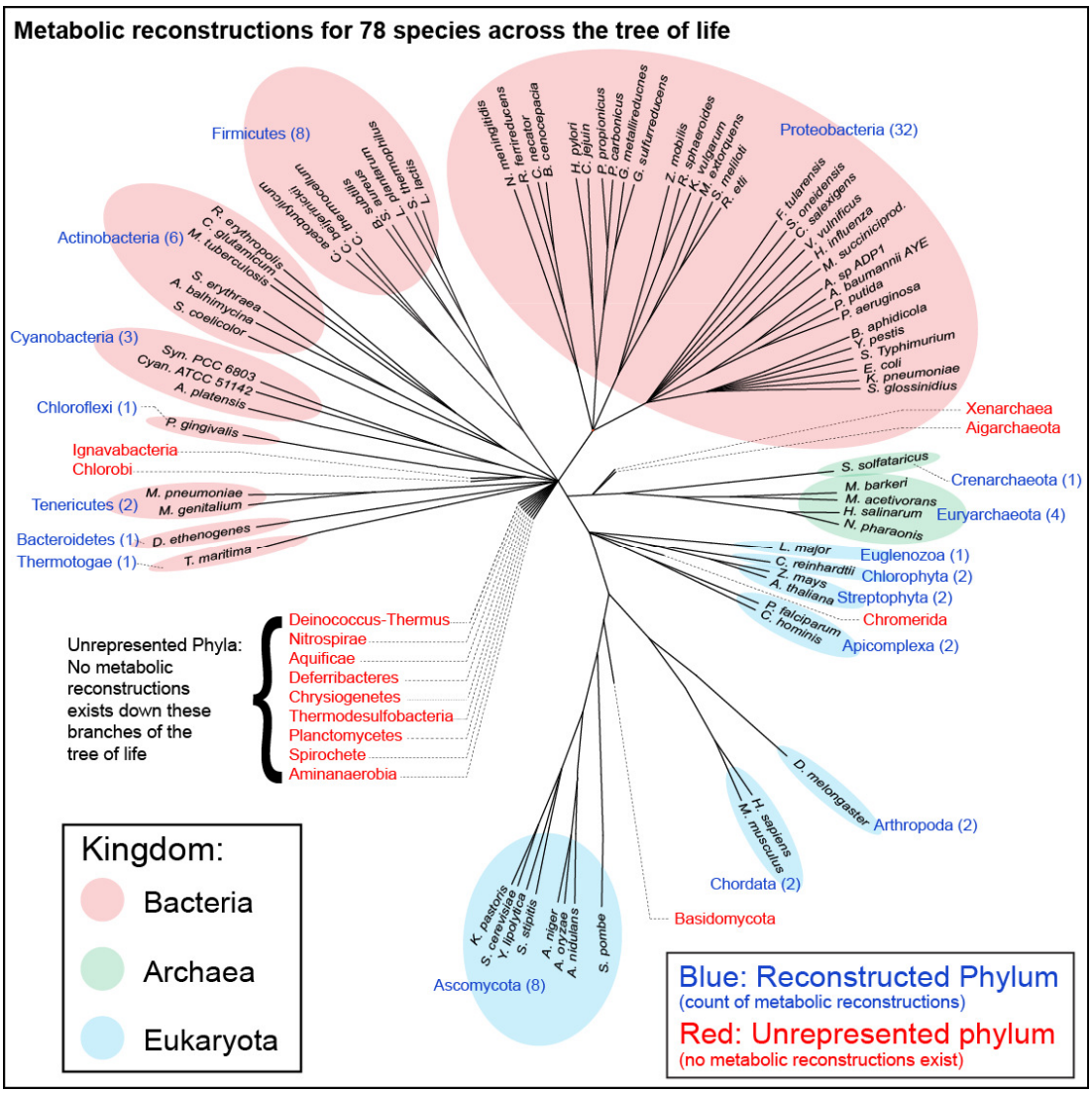


Figure 6.3 Phylogenetic coverage of GENREs.
A distribution of reconstructed species located across the phylogenetic tree of life. The Bacterial kingdom has the most organisms with reconstructed GENREs. Within the bacterial kingdom the proteobacteria phylum has the most organisms with reconstructed GENREs. In stark contrast to proteobacteria, there are many other phyla across the tree of life without any GENREs present. These are denoted as “dead-end” phyla and are colored in red. Also see the SBRG website for an interactive, up-to-date representation of reconstructed species and their location in the tree of life.

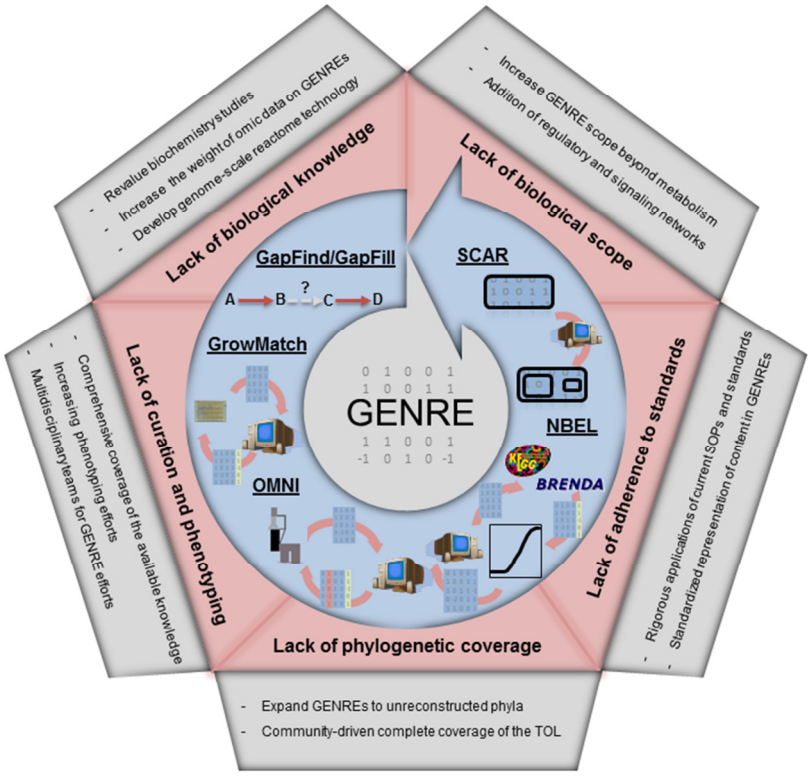


Figure 6.4. Shortcomings hampering the completeness of GENREs and main ways of improvement.
Multiple limitations currently hamper the full development of the reconstruction field (in red). Although each shortcoming can be mitigated with specific actions (in grey), the complete development of the field requires a broad and synergistic strategy to address these shortcomings as a whole. Additionally, the reconstruction and analysis process represents a unique opportunity to systematize the completion for a given GENRE in an iterative manner. Several constraint-based methods such as GapFind/GapFill, GrowMach, OMNI etc (in blue) can be applied toward GENRE refinement. The full collection of these methods have been exhaustively reviewed elsewhere.

Chapter 7 A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism for 2016

7.1 Abstract

The development of the bottom-up approach to systems biology has been driven by a genome-scale reconstruction of *Escherichia coli* K-12 MG1655 metabolism. A new version of this network is represented, named *i*ML1502, which accounts for 1502 genes, 2753 metabolic reactions, and 1202 unique metabolites. The *i*ML1502 model was first validated against growth conditions using experimental screens of 1502 gene knockout strains in 4 different conditions and achieved a greater than 90% success rate. Second, used for analysis of false-negative growth prediction that illuminates cases where alternative pathways and isozymes are yet to be discovered; Third, compared to outcomes of adaptive laboratory evolution studies to elucidate the fundamental nature of the elusive maintenance coefficients, and; Fourth, customized for specific classes of applications: 1) for conserved metabolic capabilities of *E. coli* as a species using genome sequences of over 1000 *E. coli* and *Shigella* strains; 2) for containing only primary isozyme activities that can be used to examine pathway usage under common growth and bioprocessing conditions; 3) for reactions that generate reactive oxygen species (ROS) for study of damage and repair pathways, and; 4) inclusion of 3D protein structural information of proteins. All customized models are disseminated through BiGG Models, and they enable wide applications *E. coli* metabolic systems biology, ranging from studies of protein promiscuity and underground metabolism, to pathogenesis and metabolic engineering, to evolution and phylogenetics.

7.2 Introduction

Genome-scale network reconstructions of metabolism form a common denominator for bottom-up systems biology studies (Bordbar et al., 2014). A network reconstruction represents a biochemically, genetically, and genomically (BiGG)

structured knowledge-base that contains detailed information about the target organism in a structured format (O'Brien et al., 2015). For most effective use, these reconstructions must be of high-coverage and of high quality (Monk et al., 2014b, Heavner and Price, 2015, Ebrahim et al., 2015, Ravikrishnan and Raman, 2015). New biochemical functions and metabolic capabilities are constantly being discovered, even for the extensively studied bacteria *E. coli*. Thus, these reconstructions must be periodically updated to account for new gene and cellular functions that continue to broaden their uses for discovery, new understanding and insights, and enablement of new applications (McCloskey et al., 2013b).

Here, we present an updated version of the *E. coli* metabolic network reconstruction. This new version, titled iML1502, includes newly characterized genes and reactions, the majority of which have been discovered since 2011 (Figure 1). It includes new structural information about protein and reactive oxygen species (ROS). iML1502 is the most complete and accurate *E. coli* metabolic reconstruction to date, and like its predecessors will likely aid in many new discoveries.

7.3 Results

7.3.1 Process for updating the reconstruction and its content

An updated and expanded metabolic network reconstruction of *E. coli* K-12 MG1655 was assembled and named *iML1502*. The reconstruction update process began with a review of all previously incorporated network content. As part of the reconstruction protocol (Thiele and Palsson, 2010b) confidence scores are assigned to all reactions in the reconstruction using evidence from functional assignments based on experimental data such as enzymatic assays, genetic interactions, sequence data, localization information and physiological and modeling evidence. The previous

reconstruction, iJO1366, had 350 '0 confidence reactions', meaning they had not undergone this quality check. Most of these reactions were added in the first *E. coli* metabolic reconstruction, iJE660 (Edwards and Palsson, 2000a). To fully quality control the reconstruction we started with an in-depth analysis of these '0-confidence' reactions. This process led to the discovery of vestigial errors, for example we found that the beta-frucofuranosidase reaction (FFSD) was added to the model without any gene association. This reaction allows the model to catabolize sucrose, a function that *E. coli* K-12 MG1655 is known not to be able to perform (Tsunekawa et al., 1992). Overall 332 of the '0 confidence reactions' were updated to confidence level 2 or above (**Supplementary Table S4**).

During the update of '0 confidence' reactions, we also identified metabolic subsystems that would benefit from further additions and refinement. One of these systems was the murein production network. This network was re-curated based on new information. Importantly, the genes *mrdA* and *ftsL* encoding DD-transpeptidases that have recently been discovered to be essential for proper cell division and murein elongation (Vollmer, 2012) were added. Previously, each of these genes was included as an isozyme in iJO1366, limiting the model's predictive capabilities related to murein synthesis. Beyond murein, there were several reactions missing in iJO1366 relating to reactions that elongate alcohols, aldehydes and fatty acids. And the substrates and products for several carboxylate oxidoreductase were missing from the model as a result. These changes will now allow model predictions for metabolite yields of these compounds, many of which are of interest to the metabolic engineering community for synthesis of advanced long-chain fuels and chemicals (Dellomonaco et al., 2011).

During the re-curation process, a total of 25 reactions present in iJO1366 were adjusted with changes to product or substrate usage or to reaction reversibility. For example, allantoate amidohydrolase was previously thought to catalyze both hydrolysis reactions in the allantoin degradation pathway (Agarwal et al., 2007) to produce (S)-Ureidoglycolate directly. However, the recent discovery of *S*-ureidoglycine aminohydrolase shows that these two functions are carried out by two different enzymes (Serventi et al., 2010). Thus, the allantoate amidohydrolase reaction (ALLTAM) was updated to produce S-ureidoglycine as a product and the conversion of S-ureidoglycine to S-ureidoglycolate catalyzed by the AllE aminohydrolase was added to the model. Changes were also made to some reaction reversibility constraints. For example, in iJO1366 the malate oxidase reaction (MOX) was erroneously set as reversible. In iML1502 the MOX reaction has been set to irreversible. Adjusting this reaction allowed previously removed reactions CAT, SPODM, SPODMpp, SUCASPtpp, SUCFUMtpp, SUCMALtpp and SUCTARTtpp to be re-included in iML1502 and their functions to be modeled.

Changes were also made to the reconstruction based on recent model-driven 'gap-filling' studies. These studies use contradictions between model predictions and experimental results to identify errors in the model that can then be experimentally investigated (Orth and Palsson, 2010a, Molina-Henares et al., 2010). Recent studies have used such approaches to identify new enzymes (Orth and Palsson, 2012), promiscuous isozyme function (Guzman et al., 2015) and corrections to model errors (Kumar and Maranas, 2009, Aziz et al., 2015a). Thus, throughout the iML1502 reconstruction process significant emphasis was placed on correcting FPs and FNs that have persisted in previous model versions. Some of these errors were corrected by

applying the high confidence changes suggested in a previous iJO1366 gap filling study (Orth and Palsson, 2012) where a gap filling algorithm, SMILEY (Reed et al., 2006a) was used to suggest model changes. Evidence was found supporting the reversibility of the two 2,5- diketo-D-gluconate reductase reactions, DKGLCNR1 (Habrych et al., 2002) and DKGLCNR2y (Yum et al., 1999).

In addition to previous gap filling attempts, the remaining FPs and FNs were critically examined. In iJO1366 an *aldA* gene knockout was incorrectly predicted as essential due to a glycolaldehyde dead end in the model following deactivation of glycoladehyde dehydrogenase GCALDD. This gap was corrected by adding a high confidence reversible reaction that converts glycoladehyde and glycine to 4-hydroxy-L-threonine by *ltaE* (21119630), thus providing an alternative pathway for consumption of this metabolite. Furthermore two model false negatives were filled for the genes *pabA* and *pabB*. These two gene products form a complex to catalyze 4-amino-4-deoxychorismate synthase (ADCS) a function required for production of 4-aminobenzoate (PABA) that leads to folate synthesis. However, it has been observed experimentally that ΔpabA and ΔpabB mutants grow in LB rich media. The model predicts cell death in this case. A literature search found that PABA can be transported into *E. coli* as PABA-glutamate (4abzglu) that can then be hydrolyzed by p-aminobenzoyl-glutamate hydrolase catalyzed by Abg (a complex of *abgA* and *abgB*) (Carter et al., 2007). Furthermore a transporter for abzglu has been discovered (Hussein et al., 1998). Thus it is likely that 4abzglu exists in LB media and the cell is able to transport 4abzglu into the cell for growth in LB media when *pabA* and/or *pabB* are knocked out.

Other gaps were also filled in the folate biosynthesis pathway. The folate biosynthesis pathway is lacking in humans and thus enzymes in this pathway are often

targeted for design of antibiotics. Thus, accurate understanding and model predictions for perturbations in this pathway are of the utmost importance. Deeper analysis of false negatives led to a new missing link between the functions of UbiC and PabC that was included in the model. These two genes catalyze essential reactions 4-aminobenzoate synthase (ADCL) and chorismate pyruvate lyase (CHRPL) in the folate biosynthesis pathway. Knockout of either of these genes was predicted to be lethal. However, this is not observed experimentally; knockouts of *pabC* and *ubiC* are viable when grown in M9-glucose minimal media (Baba et al., 2006). A literature search found evidence that each enzyme can each catalyze the other's function and that a *pabC* mutant can be rescued by overexpression of *ubiC* (Nichols and Green, 1992). The substrates for each enzyme differ only in the presence of a hydroxy group versus an amino group at position 4 of the cyclodiene ring, and the products are derived by elimination of the enol-pyruvyl moiety concomitantly with aromatization of the ring structure. Thus future gap-filling studies should focus on metabolite similarity and shared enzyme functions to further fill gaps in the metabolic network.

Following these model quality checks, an extensive literature and database search was performed across all *E. coli* K-12 MG1655 annotated genes to identify new or previously known metabolic reactions that were missing from the reconstruction. New metabolic content was added to the reconstruction based on this extensive search. The Pubmed, EcoCyc and KEGG (Kanehisa et al., 2014) databases were used for this purpose. All manual curation followed an established protocol (Thiele and Palsson, 2010b).

A number of newly discovered *E. coli* metabolic pathways and reactions were added to the network. New catabolic pathways for environmentally relevant metabolites

have recently been discovered. For example, a pathway for the consumption of sulphoquinovose (SQ, 6-deoxy-6-sulphoglucose) was recently elucidated (Denger et al., 2014) and added to the reconstruction. This pathway, termed sulphoglycolysis, is encoded by a ten-gene cluster (b3875-b3844) that yields dihydroxyacetone phosphate (DHAP), which powers energy conservation and growth of *E. coli*, and the sulphonate product 2,3-dihydroxypropane-1-sulphonate (DHPS), which is excreted. SQ is present in the photosynthetic membranes of all higher plants, mosses, ferns and algae and is expected to be produced at a rate of 10,000,000,000 tons (10 petagrams) per year (Harwood and Nicholls, 1979). Thus, its degradation by bacteria comprises a major portion of the organo-sulphur cycle in nature (Roy et al., 2003). Another newly discovered pathway, catalyzed by enzymes in the *phn* operon (b4092-b4108), is used for the degradation of alkylphosphonates (Kamat et al., 2011). Specifically, the enzymes that enable metabolism of methylphosphonate to phosphate and methane have been added to the reconstruction, allowing its use as a sole source of phosphorous. Furthermore, new pathways for the metabolism of curcumin (the active ingredient in tumeric) (Hassaninasab et al., 2011) and the pathway used to catabolize and process the signaling molecule AI2 have recently been discovered (Marques et al., 2011) and were added to the model.

The model was also updated to include metabolite damage and repair pathways. The impact of such damage on cellular metabolism and energy demands is increasingly being recognized (Linster et al., 2013). Such damage can occur due to side reactions of promiscuous reactions or by spontaneous chemical reactions. The productions of these reactions are useless or toxic and their unchecked buildup can be lethal to cells. New genetic and genomic evidence is elucidating conserved enzymes that repair metabolites

or pre-empt such damage. For example, the nicotinamide ring of NADH or NADPH can undergo spontaneous or enzymatic hydration to form the hydrates, NADHX and NADPHX (Marbaix et al., 2011). Spontaneous hydration is promoted by low pH or high temperature while the enzymatic reaction is mediated by a side activity of glyceraldehyde 3-phosphate dehydrogenase (Rafter et al., 1954). NADHX and NADPHX cannot act as electron donors or acceptors and inhibit several dehydrogenases (Yoshida and Dave, 1975) making them toxic to the cell. Both hydrates are reconverted to NADH or NADPH by an ATP- (or ADP-) dependent dehydratase. The repair of these metabolites thus represents an energy demand on the cell. This damage and repair mechanism has been included in the model along with 23 other mechanisms and their encoding genes.

Furthermore, eleven new genes of the haloacid dehalogenase (HAD)-like hydrolase superfamily were added to the model. This family of enzymes was largely uncharacterized until recent studies performed genome-wide analyses of this group of enzymes (Kuznetsova et al., 2006) and demonstrated that they possess phosphatase, betaphosphoglutamase, phosphonatase and dehalogenase activities on a broad set of substrates. Some HAD-like enzymes also perform important functions, including a heretofore missing link in riboflavin biosynthesis: the dephosphorylation of the intermediate 5-amino-6-ribitylamino-2,4(1*H*,3*H*)-pyrimidinedione 5′-phosphate (5aprbu). This function was recently discovered to be catalyzed by previously uncharacterized HAD enzymes YigB and YbjI in *E. coli* (Haase et al., 2013).

In total, 147 new genes were added to the reconstruction, while 12 genes were removed (b2311, b2045, b3835, b3380, b4301, b2874, b4395, b1773, b0736, b3803, b3807, b0323). These 12 genes were removed from the model because it was deemed

that there was insufficient evidence for their inclusion. For example *ubiX* was originally included in the GPR for the 3-octaprenyl-4-hydroxybenzoate decarboxylase (OPHBDC) reaction in ubiquinone synthesis because mutant phenotypes indicate that *ubiX* (b2311) and *ubiD* interact (Gulmezian et al., 2007). However these two proteins are not expected to form a complex, no biochemical evidence for the enzymatic function of UbiX is available (Nonet et al., 1987) and *ubiX* knockouts are viable while *ubiD* mutants are not. Thus, *ubiX* was removed from the GPR for OPHBDC.

With all of these new additions, the *iML1502* model represents a new, expanded and quality checked *E. coli* reconstruction. It contains 1502 genes, 2710 metabolic reactions, and 1202 unique metabolites. A summary of the content of *iML1502* and its predecessor, *i*JO1366, is presented in Table 1. Like *i*JO1366, *iML1502* contains a wide range of metabolic functions (Figure 1). The complete lists of reactions and metabolites in *i*JML1502 can be found in Supplementary Tables S2 and S3, with a list of all references used in Supplementary Table S4. The majority of new genes (80) added to this model have been characterized since *i*JO1366 was finalized in 2010 (Figure 1D). The fact that some references predate previous versions of the *E. coli* reconstruction does not necessarily mean that they were previously missed. Rather, as genes and reactions are often added on a pathway basis, complete functional pathways are typically fully elucidated over time from multiple sources. Thus, the citations in Figure 1 are spread out over time. The new genes mostly add new pathways and systems to the network, but a significant number of them fill gaps and orphan reactions in existing systems. A complete list of all new and removed genes, reactions, and metabolites can be found in Supplementary Table S5.

7.3.2 Updating the biomass composition and growth requirements

The “core” and “wild-type” biomass reactions of *i*JO1366 have also been updated in *iML1502*. These are reactions that drain biomass precursor compounds in experimentally determined ratios to simulate growth (Feist and Palsson, 2010b, Varma et al., 1993). Each component of a biomass reaction has the units of mmol/gDW (milli-moles per gram cell dry weight), and flux through a biomass reaction has the units of h^{-1}, and is equivalent to the exponential growth rate of the organism (Thiele and Palsson, 2010b). The “wild-type” biomass reaction contains the precursors to all the typical wild-type cellular components of *E. coli*, while the “core” biomass reaction contains the precursors only to essential components. Bis-molybdopterin guanine dinucleotide (bmocogdp) was removed from the core and wild type biomass objective functions due to consistent growth by mutants deficient in the ability to synthesize this metabolite.

Furthermore, succinyl co-A (succoa) was added to the core biomass function. *E. coli* has only one lysine synthesis pathway that consumes 1 succoa and produces 1 succinate. Succinate can be converted back to succoa by succinyl-CoA synthase. Thus, in steady state FBA simulations the model is able to produce lysine using the core biomass function without new succoa being produced. Realistically, some amount of succoa must be used by the lysine pathway and must be de novo synthesized as new biomass is produced, just like the other cofactors that mostly form conserved pools (e.g. NAD/NADH), which are included in the core biomass reaction. Since lysine is essential and requires the succoa consuming reaction in order to be produced, succoa is an essential cofactor and should be included in the core biomass function. This change has no effect on gene knockout results predictions but should lead to more accurate flux distributions predicted by the model.

Growth-associated maintenance (GAM) and non-growth-associated maintenance (NGAM) are the amounts of ATP consumed during cell growth and by non-growth associated processes such as maintenance of membrane gradients, respectively. GAM is a component of the biomass reaction, while NGAM is manifest as a lower bound on the separate ATP draining reaction "ATPM." These two parameters were recalculated for *iML1502* based on extensive new datasets from adaptive laboratory evolution (ALE) endpoint strains for *E. coli* K-12 MG1655 (single substrate evolutions, Glucose, Xylose, Glycerol, Acetate, Central Carbon KOs on Glucose). Such evolution studies represent a global minimum of maintenance energy and are thus useful for calculation of GAM. Using these studies, GAM was determined to be 51.09 mmol ATP gDW^{-1}, while NGAM was determined to be 3.15 mmol ATP gDW^{-1} h^{-1}. It should be noted that the GAM and NGAM in a specific strain biomass reaction can vary given the experimental data set from which they were calculated. As such, these values should be based on the experimental data that most closely matches the field of use for a modeling application. For the complete core and wild-type biomass reactions see Supplementary Table S6.

7.3.4 Prediction of all growth-supporting carbon, nitrogen, phosphorus, and sulfur sources

The *iML1502* reconstruction can be converted to a mathematical model to computationally examine *E. coli* metabolism. The model contains exchange reactions for 324 different compounds. It is therefore possible to use *iML1502* to predict the growth capabilities of *E. coli* on a very wide range of media conditions. As a demonstration of the prediction of growth capabilities, FBA was used to predict growth on every possible carbon, nitrogen, phosphorus, and sulfur source, one at a time, under both aerobic and

anaerobic conditions (Supplementary Table S7). Aerobically, a total of 187 carbon sources (out of 297 carbon containing compounds), 94 nitrogen sources (out of 182), 50 phosphorous sources (out of 67) and 11 sulfur sources (out of 30) were found to be growth supporting (Table 2). There are several reasons why a carbon containing metabolite cannot serve as a carbon source. First, not all extracellular compounds have transport reactions that allow them to enter the cell. Some may only have efflux reactions that allow them to be excreted. Second, some compounds are not connected to the central reactions of metabolism, from which all essential biomass components are constructed. For example, cob(I)alamin can be converted only to vitamin B12, but not to any other biomass components. Third, carbon sources must also generally serve as energy sources for *E. coli*, so a highly oxidized compound such as CO_2 cannot be growth supporting. Not all compounds can serve as nitrogen, phosphorus, and sulfur sources for similar reasons.

7.3.5 Prediction of gene essentiality

The iML1502 model can predict the effect of gene knockouts through the "gene-protein-reaction' relationship, or GPR. Every reaction in *i*ML1502 is linked to its catalyzing protein that in turn is linked to its encoding gene via this GPR relationship that enables to model to predict the effect of gene knockouts. FBA was used to predict the optimal growth rate of *E. coli* growing in four different conditions: glucose aerobic M9 minimal medium, glucose anaerobic M9 minimal medium, lactate aerobic M9 minimal medium, and succinate aerobic M9 minimal medium (Orth et al., 2011). Complete results of these screens for all 1502 genes knocked out one at a time can be found in Supplementary Table S6. When compared to experimental gene essentiality data, most of the predictions made by *i*ML1502 are correct, confirming its overall accuracy (91%).

There are incorrect growth predictions made by iML1502. Comparing computational predictions to gene essentiality data leads to four possible outcomes. Correct predictions come in two forms: True Positives (TP; model growth/experimental growth) and True Negatives (TN; model no growth/experimental no growth). False predictions are classified as False Positives (FP; model growth/experimental no growth) or False Negatives (FN; model no growth/experimental growth). False positive predictions are the results of a model possessing unrealistic capabilities, such as pathways that are not normally expressed during the particular growth conditions. Because *iML1502* is a metabolic network model that does not contain regulatory systems, FP predictions are possible. FN cases, on the other hand, indicate that some realistic content such as an essential transport or enzymatic reaction may be missing from the model. Both situations offer opportunities for model improvement and these predictions can be used to drive model-based biological discovery (Reed et al., 2006a).

As a result of efforts to remediate previous model errors, iML1502 showed a 10.6% increase over iJO1366 in the models' ability to predict gene essentiality against the complete Keio Collection grown on glucose M9 minimal media in aerobic conditions, as determined by the Matthews Correlation Coefficient (MCC). This included a reduction in the number of FPs and FNs by 2 and 15, respectively. Improvement was seen across all other conditions as well. It was predicted that the remaining 38 false positives were likely due to the regulatory conditions specific to E. coli K-12 MG1655 growth on glucose aerobic M9 minimal media. These were considered further when construction iML1502-iso (see Modularized extensions of *iML1502*).

7.3.6 Customization of iML1502 to enable broad use cases

Comprehensive genome-scale models of metabolism can be customized them to specific situations. Such customization is achieved through adding or removing content from the full reconstruction to describe specific conditions of interest. Below we detail four specific modules that can be used to extend the predictive capabilities of iML1502 to address new questions. First, the reconstruction can be reduced to contain only the metabolic content shared among all sequenced *E. coli* strains to examine the capabilities of the so-called core genome (Lukjancenko et al., 2010, Medini et al., 2005). Then strain specific capabilities can be determined as augmentations to this core. Second, iML1502 can be expanded to include the reactions (many of which are spontaneous non-catalyzed reaction) that generate reactive oxygen species (ROS) as by-products. Each customized model is released with this paper and housed in BiGG Models.

1) iML1502-csvd: A large number of strains of a given bacterium are now being sequenced. This opens up the possibility to study and define a bacterial 'species,' its common functionalities and its variability. Over a 1000 genomes of diverse *E. coli* strains have now been sequenced. Based on these sequences it is possible to establish the conserved metabolic capabilities of *E. coli* as a species by determining which metabolic genes are shared among all strains (Figure 2A).

We compared metabolic genes in iML1502 across 1,129 sequenced strains of *E. coli* and *Shigella* . We found that 983 metabolic genes were shared among >99% of the strains. iML1502 was stripped of those genes not present in greater than 1111 strains (99% of strains) to form iML1502-csvd. This model of conserved *E. coli* metabolism contained 983 genes, 1867 reactions and 1202 unique metabolites making it similar in size to the conserved model formed among 55 strains of *E. coli* (Monk et al., 2013).

iML1502-cons is auxotrophic for nutrients including L-proline, L-phenylalanine, L-tryptophan, L-arginine, L-tyrosine, L-glutamate, L-glutamine, biotin, thiamine and tetrahydrofolate indicating that the ability to synthesize these molecules is not conserved across all strains of *E. coli*. When the *in silico* minimal media is supplemented with these nutrients, the conserved metabolic model is able to grow on 114/187 C sources, 41/94 N sources, 6/11 S sources and 41/50 P sources. iML1502-csvd is provided with this manuscript for use in applications studying conserved *E. coli* metabolic functions.

2) iML1502-iso: Each prior iteration of the *E. coli* K-12 MG1655 metabolic reconstruction generated a marked improvement in prediction of gene essentiality. This improvement mostly stems from the improvement in the number of false negatives predicted as additional network content is added to the model. Conversely, there has been little improvement in the number of false positives. This lack of improvement can largely be attributed to the inclusion of secondary isozymes to the GPRs of metabolic reactions. While most of these are correct, the window of observation for gene essentiality experiments is often not long enough to observe the up-regulation of alternate isozymes in KO strains. Considering that many of these isozymes take multiple days to be effectively expressed and utilized they can be eliminated from iML1502 when it is used for sort term predictions. By contextualizing iML1502 to short term growth on glucose aerobic M9 minimal media, iML1502-iso was able to predict gene essentiality data with an MCC of 0.83 by reducing the number of FPs from 38 to 16. iML1502-iso is customized for use to interpret or design experiments performed in glucose aerobic M9 minimal media in which the window of observation is short enough so that up-regulation of reactions catalyzed by secondary isozymes will not take place.

The iML1502-iso model was constructed by filtering out genes with low RNAseq expression compared to the primary enzyme catalyzing the reaction. This identified 152 secondary isozymes that resulted in 194 reaction GPR changes and the complete removal of 86 gene products from the model. The 86 genes that were removed synthesize proteins that have no primary function detailed in the model under these specific conditions. A set of additional genes and reactions was also removed due to known roles in solely anaerobic conditions. These reactions included OXAMC and CBMKr that are used under anaerobic conditions to metabolize allantoin as a nitrogen source. Furthermore, there were several gene knockouts that incorrectly predicted cell growth due to the model availability of lowly expressed alternative pathways/reaction. Additionally, the XAD reactions allowed the models to …. and were removed due to very low (<20 counts) expression on glucose aerobic M9 minimal media.

In order for *E. coli* to grow in this specific condition, specific processes or pathways must be active in order for the cell to grow. One such change was made to facilitate the acquisition of iron (III) by adding enterobactin to the core and wild-type biomass objective functions. In several growth conditions (i.e. LB, anaerobic conditions, MOPS media) iron (II) is present in low levels in the media, which is the soluble oxidative state of iron that can taken up directly through a variety of direct transport systems. In glucose aerobic M9 minimal media, however, most all iron will exist in it insoluble, oxidized iron (III) state. In this case, a siderophore, primarily enterobactin in *E. coli* K-12 MG1655, as well as its transport proteins are required to effectively import iron into the cell for use as a cofactor. Adding this compound to the biomass objective function will require the pathways producing and transporting enterobactin to be active in order to grow in these conditions.

3) iML1502-ROS: As genome-scale reconstructions expand in scope, modeling of some capabilities may only be desired in certain situations. One such situation describing ROS production and detoxification in *E. coli* (Brynildsen et al., 2013a). However since this study appeared, the metabolic network has been expanded by over 200 new genes and reactions. Therefore we updated the description of ROS production using iML1502. We added 334 ROS production reactions to the model based on identification of 167 ROS production sources (an increase of 34 reactions over iAF1260-ros (Brynildsen et al., 2013a)) based on an analysis of enzymes that utilize reduced flavin, quinol and transition metal functional groups which have been shown to be a source of O_2^- and H_2O_2 (Massey, 1994, Messner and Imlay, 1999). Incorporating these reactions into the model allows for simulation of ROS production and mitigation in *E. coli*. We found that knockout of 41 different genes reliably increased ROS production, in line with previous studies that have experimentally validated this approach. The iML1502-ROS model is provided here for use in studying ROS production and detoxification.

4) iML1502-PRO: Integrating genome-scale models with three-dimensional protein structures (GEM-PRO) has proven to be a powerful approach to extend the scope and predictive capability of genome-scale models. GEM-PRO enables studies that address how properties of individual proteins impact an entire metabolic network. A previous version of the GEM-PRO for *E. coli* K-12 MG1655 has been released.

This GEM-PRO, termed iRC1366-GP, contained 1365 genes, of which 98 had no structural information, 465 had available crystallographic structures and 803 were homology modeled. Now, with iML1502, 153 new metabolic genes have been incorporated into the GEM-PRO model ready for structural modelling (12 genes have been removed). 97% of these genes had crystal structures available in PDB, in which

the majority were high resolution (< 3Å, see Supplementary Information) deemed to be of sufficient quality. The remaining structures were obtained using a well-established homology modelling protocol (Zhang, 2008, Roy et al., 2010). The iML1502-pro model can be used in structural applications including prediction of behavior at different temperatures (Chang et al., 2013a), docking predictions for antibiotic design (Chang et al., 2013c) and discovery of new promiscuous enzyme functions (Guzman et al., 2015).

Connecting each enzyme to its three-dimensional structure also allows for a finer characterization of the gene-protein-reaction relationship. This relationship is the link from genotype to phenotype in a genome-scale model, it allows for connection between gene coding (G) for a protein (P) that then catalyzes a reaction (R) used in the cell. With three dimensional structures of proteins we can now get a finer grained detail into the catalytic process by actually determining the catalytic domain that performs an enzymatic transformation. This enables a new relationship to be formed, termed the "DGPR" or domain-gene-protein-reaction. The GEM-PRO enables us to assign all catalytic domains in proteins within the nextwork as well as their coding regions in a specific gene.

7.4 Discussion

iML1502 is the highest quality, most comprehensive metabolic reconstruction of *E. coli* K-12 MG1655 to date. It was established here by validating the model's predictions across mutant phenotype screens for all genes in the model across four conditions different conditions. By curating all of the low confidence reactions in the previous model, the average confidence of reactions increased from 2.58 to 3.25. This process also illuminated subsystems in the model that could benefit from further curation. New metabolic content has been added to iML1502, including the newly

elucidated pathways for metabolism of environmentally relevant metabolites such as sulphoquinovose, methylphosphonate, curcumin and AI2. Furthermore, a significant effort was made to gap fill and correct known model errors. When comparing the essentiality predictions of iML1502 to the previous model a 10% increase in predictive accuracy is observed across all gene knockouts and growth conditions.

The presented reconstruction provides a useful tool for predicting the metabolic state of *E. coli* K-12 MG1655 as well as a high quality scaffold for constructing the next generation of cellular models. For the past decade, previous versions of this reconstruction have been used in a wide range of studies, from the discovery and characterization of new metabolic genes (Orth & Palsson, 2010; Reed et al, 2006b) to the design of high-yield production strains for industrially valuable compounds (Feist et al, 2010; Kim et al, 2008). Like its predecessors, iML1502 is expected to have many practical applications (Feist & Palsson, 2008). Given that this strain of E. coli is perhaps the most exhaustively studied prokaryote, this metabolic reconstruction can also act as an ideal platform for constructing the next generation of cellular models. Some of these have already come to fruition in the form of models of metabolism and expression as well as models with integrated protein structural information. The predictability of these new models, however, hinges on the quality of the scaffold, as any errors may percolate through the model extensions and invalidate results.

In addition to a metabolic reconstruction specific to only the *E. coli* K-12 MG1655 strain, we present four specific modules that can be used to augment iML1502's capabilities and simulation metabolic functions under unique situations: iML1502-cons, iML1502-iso, iML1502-ros and iML1502-pro. The iML1502 genome-scale metabolic network reconstruction of *E. coli* is the latest update to one of the workhorse models of

the microbial systems biology community. Some of the applications, such as prediction of growth phenotypes in different media and with gene knockouts, have been presented here. The accuracy of the model has been confirmed by comparisons to experimental data. More advanced uses of *iMl1502* include guiding the discovery of metabolic genes and reactions and design of metabolic engineering production strains. This model can now be integrated with genome-scale network reconstruction of other cellular systems such as transcriptional regulation and transcription and translation. *iML1502* is the most advanced and comprehensive metabolic reconstruction of any microorganism to date, and can thus continue to serve as a basis for the metabolic reconstruction of other bacteria. Based on the success of it its predecessors, one may expect that *iML1502* will be an important tool in microbial systems biology for years to come.

7.5 Materials and Methods

7.5.1 Network reconstruction procedure

The *iML1502* reconstruction was assembled by updating the *iJO1366 E. coli* metabolic reconstruction (Orth et al., 2011). A 96-step procedure (Thiele and Palsson, 2010b) for metabolic network reconstruction was followed when adding new genes, reactions, and metabolites to form *iML1502*. The reconstruction was assembled using the SimPheny (Genomatica Inc., San Diego, CA) software platform. All new metabolites were checked against public databases (e.g. KEGG, PubChem) for correct structure and charge at a pH of 7.2. New reactions were mass and charge balanced and reversibility was assigned based on experimental studies, thermodynamic information, or the heuristic rules in the standard reconstruction protocol (Thiele and Palsson, 2010b). Reactions were associated with genes and functional proteins to form GPRs. The *iML1502* model was exported from SimPheny as an SBML file and the COBRApy

Toolbox was used for additional model testing. The Gurobi linear programming solver was used for all optimization procedures.

7.5.2 In vivo phenotypic screens

The *iML1502 E. coli* K-12 MG1655 metabolic reconstruction was used to make computational predictions. The parent strain of the Keio Collection, BW25113, is derived from K-12 MG1655 and is missing several genes that are present in K-12 MG1655: *araBAD*, *rhaBAD*, and *lacZ*. Therefore, flux through the associated reactions without isozymes (ARAI, RBK_L1, RMPA, LYXI, RMI, RMK, and LACZ) was constrained by setting the upper and lower flux bounds of the reactions to zero. The lower bounds of exchange reactions were set to default values to simulate minimal media. For aerobic growth, oxygen uptake was allowed by setting the lower bound of the oxygen exchange reaction to -18.5 mmol gDW^{-1} h^{-1}. Anaerobic growth was modeled by setting the lower bound of this reaction to zero. For growth on glucose, the lower bound of the glucose exchange reaction was set to -8 mmol gDW^{-1} h^{-1}. For growth on L-lactate and succinate, the lower bounds were set to -16 mmol gDW^{-1} h^{-1}. After setting the bounds for each condition, the predicted effect of the single deletion of each gene in *iML1502* for each condition was computed using the COBRApy Toolbox delete_model_genes function, which uses GPRs to constrain the appropriate reactions to carry zero flux and then predicts maximum growth using FBA. A gene was considered essential for the simulated condition if deletion of the gene reduced optimal growth rate to less than 5% of the wild-type strain as computed by FBA.

Knockout strains were taken from the Keio Collection (Baba *et al*, 2006; Yamamoto *et al*, 2009) (supplied by Open Biosystems), a genome-scale collection of *E. coli* K-12 gene knockouts created by the method of Datsenko and Wanner (Datsenko &

Wanner, 2000). Stocks of knockout mutants were streaked onto LB agar with kanamycin (50 μg/mL) from odd-numbered Keio collection plates only. Two single colonies of the Keio knockout strain were inoculated into 96-well plates containing 200 μL of LB media with kanamycin (50 μg/mL) and incubated overnight at 37°C without shaking. Plates were then centrifuged and pelleted cells were washed twice with 200 μL of 1X M9 salts per well. Disposable replicator pins were used to transfer cells from the pre-culture plate to four new plates, two containing glucose M9 minimal medium, one containing lactate M9 minimal medium, and one containing succinate M9 minimal medium. There was 200 μL of minimal medium per well, and the minimal media also contained 50 μg/mL kanamycin. The four 96-well plates were covered with Aeraseal breathable film (Sigma) to minimize cross-well contamination. For aerobic conditions, the plates were incubated at 37° C without shaking in a sterile cabinet. For anaerobic conditions, the plates were incubated at 37° C in an anaerobic chamber ($[O_2]$ < 50 ppm). After 48 h, absorbance at 600 nm of each well was determined using a VERSAmax microplate reader (Molecular Devices, Sunnyvale, CA).

A well was considered to have no growth or slow growth if its absorbance was less than a cut-off value specific to each of the four conditions. The cut-off value was determined by visual inspection of a histogram of absorbance values for all wells measured from a given condition. “Normal” growers are supposed to result in measurements lying in the roughly Gaussian shaped distribution that makes up most of the data, while slow- and non-growers are supposed to result in measurements falling outside the upper 0.95 area of the Gaussian distribution. The cut-offs were OD_{600} = 0.26 for glucose aerobic plates, OD_{600} = 0.21 for glucose anaerobic plates, OD_{600} = 0.21 for lactate aerobic plates, and OD_{600} = 0.20 for succinate aerobic plates. If both colonies of

a gene knockout were determined to have normal growth, then the gene was considered a true positive (TP) if the model had predicted growth for the knockout, or a tentative false negative (TFN) if the model indicated the knockout should not have grown or previous experiments in glucose MOPS minimal medium had indicated the gene was essential (Baba *et al*, 2006). Similarly, if both colonies of a gene knockout were determined to have slow/no growth, then the gene was considered a true negative (TN) if the mode predicted the same outcome, or a tentative false positive (TFP) otherwise. A gene was considered inconclusive (INC) if for any condition only one colony showed slow/no growth.

A second round of screening was performed for TFN, TFP, and INC gene knockouts. TFP and INC gene knockouts were rescreened with two colonies each; TFN gene knockouts were rescreened with four colonies and their pellets were washed four times before transfer to minimal media instead of twice to ensure that growth was not due to contamination from trace amounts of rich media from the precultures. For TFP and INC gene knockouts, a gene was considered essential for a condition if at least one colony showed slow/no growth in the condition in both the first and second round screens. If an essential gene was predicted by the model to be non-essential in the experimental condition, then the gene was concluded to be a genuine false positive (FP) for that condition. For TFN gene knockouts, a gene was considered a genuine false negative (FN) if two of four colonies demonstrated normal growth in the secondary screen.

7.5.3 Constraint-based modeling

The *iML1502* model, constructed in SimPheny, was exported as an SBML file and used to perform simulations and constraint-based analyses using the COBRApy

Toolbox and Gurobi linear programming solver. The constraint-based model consists of a stoichiometric matrix (**S**) with *m* rows and *n* columns, where *m* is the number of distinct metabolites and *n* is the number of reactions. Each of the *n* reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 mmol gDW^{-1} h^{-1} and a lower bound of -1000 mmol gDW^{-1} h^{-1}, while irreversible reactions have a lower bound of zero. FBA can be used to identify optimal steady-state flux distributions of constraint-based models. Linear programming is used to find a solution to the equation **Sv** = 0, given the set of upper and lower bounds and an objective function defined by a vector **c** of length *n*. **v** is a vector of reaction fluxes of length *n*. Typically, **c** is a vector of 0s with a 1 at the position of the reaction flux to be maximized or minimized. For a thorough description of FBA, see (Orth et al, 2010).

For most growth simulations, the core biomass reaction is set as the objective to be maximized. Certain reactions that are not used under typical growth states are by default constrained to carry zero flux. These reactions are CAT, DHPTDNR, DHPTDNRN, FHL. The NGAM constraint is imposed by a lower bound of 3.15 mmol gDW^{-1} h^{-1} on the reaction ATPM. The exchange reactions that allow for extracellular metabolites to pass in and out of the system are defined such that a positive flux indicates flow out. All exchange reactions have a lower bound of zero except for glucose (-10 mmol gDW^{-1} h^{-1}) and oxygen and all inorganic ions required by the biomass reaction (-1000 mmol gDW^{-1} h^{-1}).

7.5.4 Prediction of different carbon, nitrogen, phosphorus, and sulfur sources

The possible growth-supporting carbon, nitrogen, phosphorus, and sulfur sources of *E. coli* were identified using FBA. First, all exchange reactions for extracellular metabolites containing the four elements were identified from the metabolite formulas.

Every extracellular compound containing carbon was considered a potential carbon source, for example. Next, to determine possible growth supporting carbon sources, the lower bound of the glucose exchange reaction was constrained to zero. Then the lower bound of each carbon exchange reaction was set, one at a time, to -10 mmol gDW^{-1} h^{-1}, and growth was maximized by FBA using the core biomass reaction. The target substrate was considered growth supporting if the predicted growth rate was above zero. While identifying carbon sources, the default nitrogen, phosphorus, and sulfur sources were ammonium (nh4), inorganic phosphate (pi), and inorganic sulfate (so4). Prediction of growth supporting sources of these other three elements was performed in the same manner as growth on carbon, with glucose as the default carbon source.

7.5.5 Gene essentiality predictions

To simulate the effects of gene knockouts, the *iML1502* model with its default constraints and core biomass reaction objective was modified to match the genotype of *E. coli* BW25113 (see ***In vivo* phenotypic screens**). For growth on glucose, the lower bound of the glucose exchange reaction was set to -10 mmol gDW^{-1} h^{-1}. For growth on glycerol, the lower bound of the glucose exchange reaction was set to zero while the lower bound of the glycerol exchange reaction was set to -10 mmol gDW^{-1} h^{-1}. All 1362 genes in the model were knocked out one a time and growth was simulated by FBA using the delete_model_genes COBRApy Toolbox function. Gene knockout strains with a growth rate above zero were considered non-essential. The newly identified essential genes were added to the lists of essential genes under both conditions, while the genes whose essentiality was identified as uncertain were not changed from their original designations.

7.5.6 Mapping to other *E. coli* strains

The protein sequences of all available *E. coli* and *Shigella* strains (1129 strains total) were downloaded from the RAST database (CITE RAST). The RAST server get_corresponding_genes function was used to identify orthologs between E. coli K-12 MG1655 and each of the other strains. Genes were considered conserved if they were present in another organism at greater than 80% percentage identity with a best bidirectional hit (BBH). Those genes that were not shared at this cutoff in less than 99% of strains were then removed from the iML1502 model to form iML1502-consv. This model was investigated to determine which components of the biomass function could not be produced. When a strain is unable to synthesize a certain biomass component, it either has an alternate route to produce this biomass component or an auxotrophy requiring transport of this metabolite to sustain growth. This method can accurately determine known strain-specific auxotrophies.

To simulate growth on all possible C, N, P and S sources the model was provided with exchange reactions for those components that it was auxotrophic for (btn, glu-L, dttp, thmpp, 2ohph, trp-L, 10fthf, udcpdp, tyr-L, phe-L) with a lower bound of -1. The biomass function was modified by removing four components that could not be exchanged (pe161_p, pe160_p, kdo2lipid4_e, murein5px4p_p) to form a conserved biomass function: 'Ec_biomass_iML1502_CONS_53p95M'. This biomass function was used to optimize for growth of the model by individually replacing the sole source of C, P, N or S. Growth was considered to be present if the predicted growth rate was 5% greater than the growth rate with no exogenous source of C, N, P, or S (excluding the exchange reactions provided auxotrophies).

7.6 Acknowledgements

Chapter 7, in part, is a reprint of the material Monk JM*, Lloyd CJ*, Brunk L, Mih N, King Z, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism for 2016. *In Preparation*. The disseration author was the primary author (equally contributing with Colton Lloyd) of this paper.

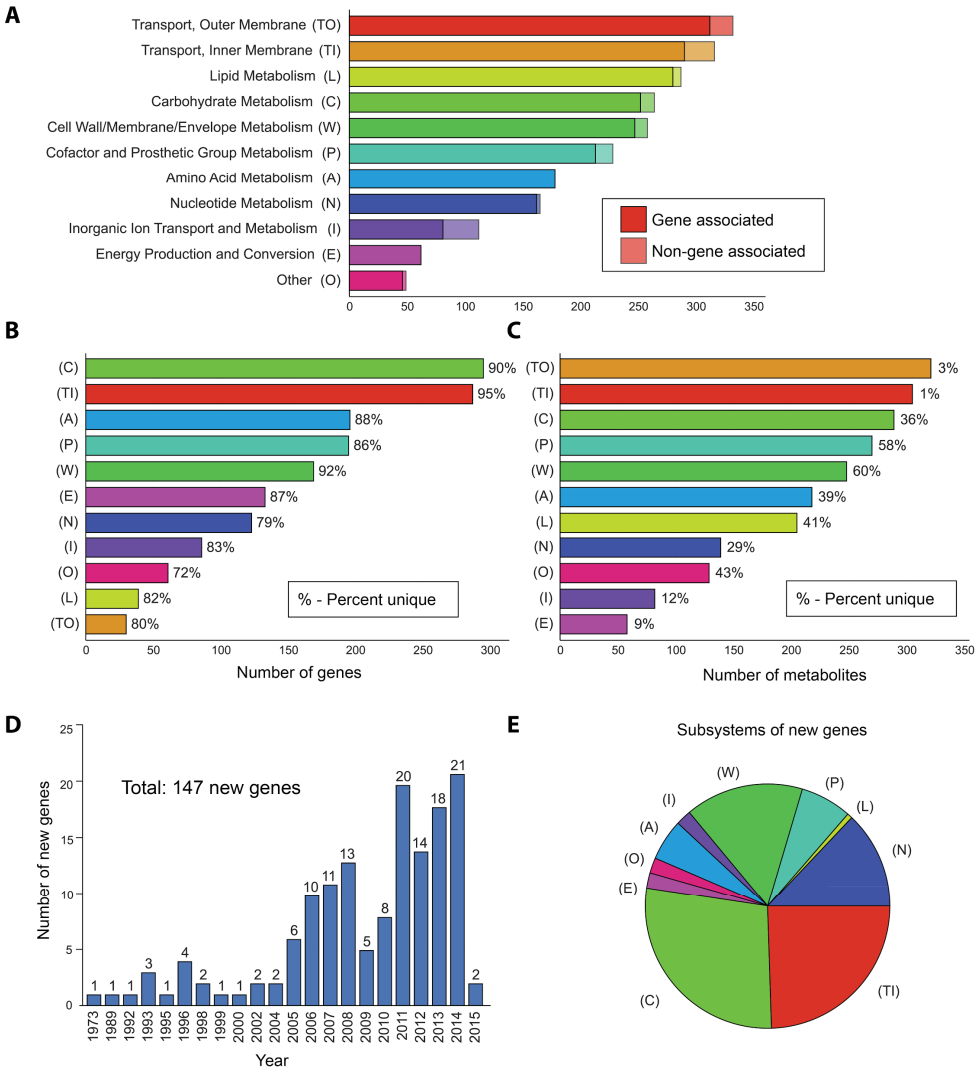
Table 7.1. Properties of *i*ML1502 and *i*JO1366

	***i*ML1502 (this study)**	***i*JO1366 (Orth *et al*, 2011)**
Included genes	1504 (35%)	1367 (32%)
Unique functional proteins	1380	1254
Multigene complexes	284	262
Genes involved in complexes	508	475
Instances of isozymes	438	365
Reactions	2372	2251
Biochemical evidence	1763	1343
Genetic evidence	378	283
Physiological evidence	332	277
Sequence evidence	106	83
Metabolic reactions	1575	1473
Cytoplasmic	1319	1272
Periplasmic	241	193
Extracellular	15	8
Transport reactions	797	778
Cytoplasm to periplasm	460	447
Periplasm to extracellular	334	329
Cytoplasm to extracellular	3	2
Gene-protein-reaction associations		
Gene association (metabolic/transport)	1510/698	1382/706
Spontaneous/diffusion reactions	19/25	21/14
Total (gene associated and no association needed)	1529/723 (95%)	1403/720 (94%)
No gene association (metabolic/transport)	46/74 (5%)	77/58 (6%)
Exchange reactions	338	330
Metabolites		
Unique metabolites	1202	1136
Cytoplasmic	1103	1039
Periplasmic	462	442
Extracellular	338	324

Table 7.2. Gene essentiality predictions on 4 minimal medias.

	Experimental	
	Growth	**No Growth**
	On Glucose	
Model Growth	931	41
Model No Growth	10	83
	On Glucose - Anaerobic	
Model Growth	922	42
Model No Growth	15	86
	On Lactate	
Model Growth	916	48
Model No Growth	9	92
	On Succinate	
Model Growth	909	51
Model No Growth	9	96

Figure 7.1. Distribution of the reactions, genes, and metabolites in iML1502 by functional category.
(**A**) The number of reactions in each of eleven categories. Each reaction was assigned to one of 36 subsystems during the reconstruction process, and these subsystems were then assigned to broader categories. Non-gene-associated (orphan) reactions are indicated by the lighter portion at the far right of each bar. (**B**) The number of genes with associated reactions in each category. The number of genes unique to each category (i.e., associated only with genes in one category) is given as a percentage. (**C**) The number of unique metabolites that participate in at least one reaction in each category, with the number of metabolites unique to each category indicated. (**D**) Histogram of the years in which the function of each new gene was first unambiguously identified. Most new genes were characterized after iJO1366 was published in 2011, while some were characterized previously but not included in the E. coli metabolic reconstruction (see text). (**E**) Classification of each of the 147 new genes in iML1502 by metabolic subsystem.


A
Transport, Outer Membrane (TO)
Transport, Inner Membrane (TI)
Lipid Metabolism (L)
Carbohydrate Metabolism (C)
Cell Wall/Membrane/Envelope Metabolism (W)
Cofactor and Prosthetic Group Metabolism (P)
Amino Acid Metabolism (A)
Nucleotide Metabolism (N)
Inorganic Ion Transport and Metabolism (I)
Energy Production and Conversion (E)
Other (O)
Gene associated
Non-gene associated
0 50 100 150 200 250 300 350
B
(C) 90%
(TI) 95%
(A) 88%
(P) 86%
(W) 92%
(E) 87%
(N) 79%
(I) 83%
(O) 72%
(L) 82%
(TO) 80%
% - Percent unique
0 50 100 150 200 250 300
Number of genes
C
(TO) 3%
(TI) 1%
(C) 36%
(P) 58%
(W) 60%
(A) 39%
(L) 41%
(N) 29%
(O) 43%
(I) 12%
(E) 9%
% - Percent unique
0 50 100 150 200 250 300 350
Number of metabolites
D
Total: 147 new genes
Number of new genes
0 5 10 15 20 25
1 1 1 3 1 4 2 1 1 2 2 6 10 11 13 5 8 20 14 18 21 2
1973 1989 1992 1993 1995 1996 1998 1999 2000 2002 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Year
E
Subsystems of new genes
(W)
(P)
(L)
(N)
(TI)
(C)
(E)
(O)
(A)
(I)

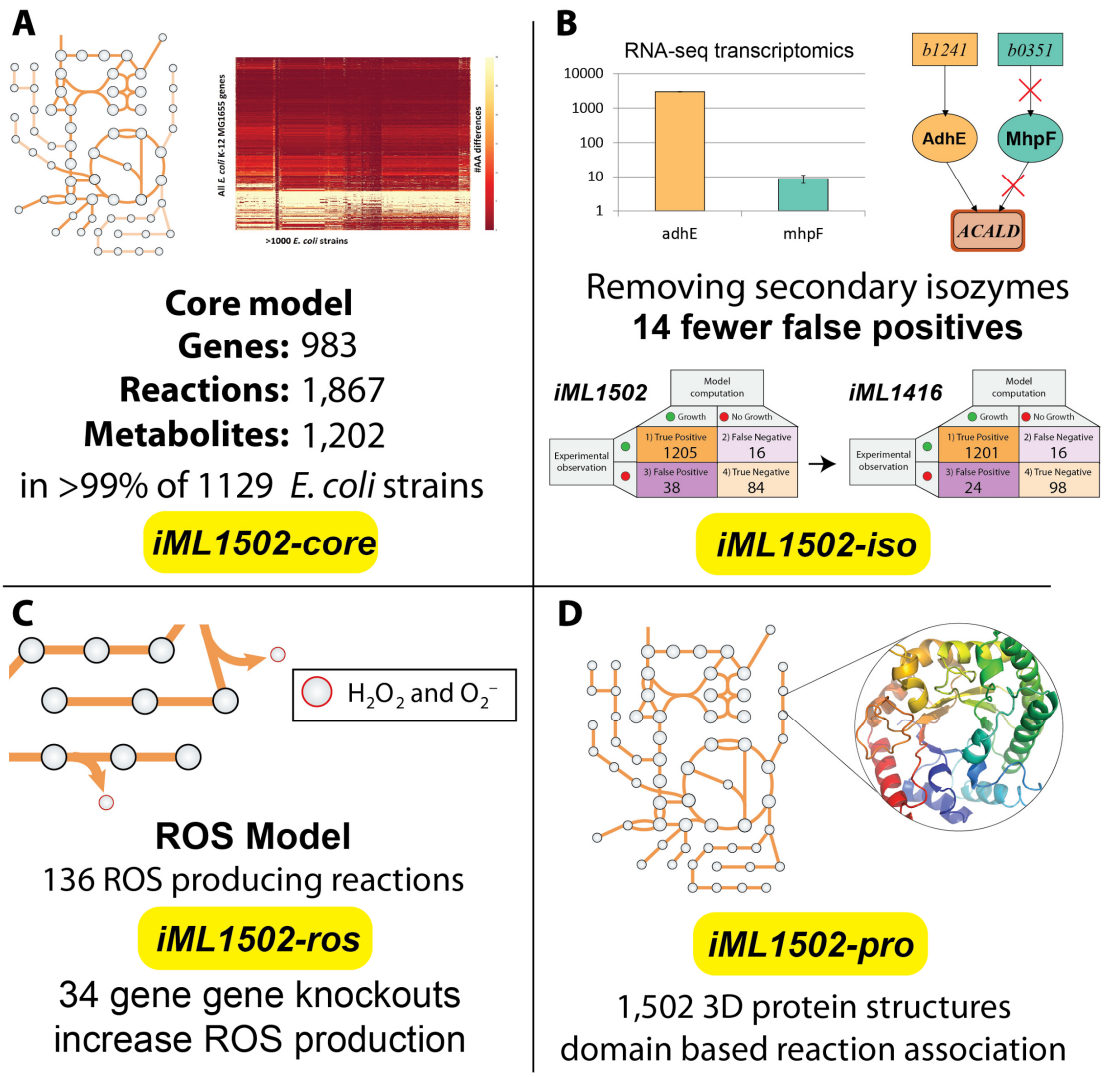


Figure 7.2. Extensions of the iML1502 model.
A) the content of iML1502 was compared to over 1000 sequenced E. coli strains. 983 of the genes are present in more than 99% of these strains. These genes form a core model. B) Model predictions were improved by detailing isozyme usage in E. coli. C) ROS production by E. coli was added to the model. D) The model was combined with 3D protein structures to.

Chapter 8 A genome-scale reconstruction of metabolism and associated protein structures in *Staphlyococcus aureus* USA300

8.1 Abstract

Staphylococcus aureus is a gram-positive pathogen of humans and animals, causing significant morbidity, mortality, and economic loss worldwide. USA300, a clone of methicillin-resistant *Staphylococcus aureus* (MRSA), is a major source of community-acquired infections in the USA, Canada, and Europe. A genome-scale model (GEM) of metabolism in this facultative anaerobic opportunistic pathogen was built based on current genomic data, literature, and physiological information to elucidate the metabolic underpinnings of this strain's distinctive epidemiological and virulence properties. The GEM comprises 1399 metabolic processes representing approximately 29% of all protein-coding regions. The GEM was extensively validated against experimental observations and correctly predicts the main physiological properties of the wild-type strain, such as aerobic and anaerobic respiration and fermentation. These results indicate that the GEM is useful in assisting future experiments to elucidate the interrelationship of bacterial metabolism and antibiotic resistance. To help directing future studies for novel chemotherapeutic targets, we conducted a large-scale *in silico* gene deletion study that identified 102 metabolic reactions essential to this bacteria's survival. The reason for the overwhelming success of the USA300 clone is not known, thus the GEM-PRO presented here should be of use for further integrated studies into the genetic factors that play a role in the success and virulence of this strain.

8.2 Introduction

Staphylococcus aureus strain USA300 is a strain of gram-positive bacteria responsible for Methicillin-resistant *Staphylococcus aureus* (MRSA) that has emerged as the predominant cause of community-associated infections in the United States, Canada and Europe (Moran et al., 2006). Today in the United States more deaths are attributed

to MRSA infections than to HIV/AIDS (Bancroft, 2007, Klevens et al., 2007). USA300 was first isolated in September, 2000, and has been implicated in wide-ranging and epidemiologically unassociated outbreaks of skin and soft tissue infections in healthy individuals (Diep et al., 2006). In 2006 the CDC reported that 64% of MRSA isolated from infected patients were of the USA300 strain, an increase from 11.3% since 2002 (Tattevin et al., 2009) indicating a rapid spread throughout the country. USA300 is capable of producing rapidly-progressing fatal conditions in humans that causes a wide variety of diseases, ranging from superficial skin and soft tissue infections to life-threatening septicaemia, endocarditis, and toxic shock syndrome. USA300 alone causes an estimated 20-40 thousand annual deaths worldwide (Highlander et al., 2007).

Although many of the mechanisms for antibiotic resistance and infection have been elucidated for this organism, there is little published information regarding the basic and systemic biochemical function of *S. aureus* USA300, especially under carefully controlled environmental conditions in chemically defined media. The annotated genome of a microorganism, in conjunction with biochemical, physiological and 3D structural data, can be used to reconstruct a metabolic network integrated with protein structures for that organism (Chang et al., 2013a). Such reconstructed networks consist of a set of chemical reactions that together comprise the known metabolic transformations that take place in a particular organism linked to the 3D structure of the enzymes catalyzing these reactions.

Genome-scale models (GEMs) combined with protein structure (GEM-PRO) represent a biochemically and genetically structured database that can be used to predict and simulated the metabolic features of an organism (Bordbar et al., 2014, O'Brien et al., 2015). With the imposition of appropriate constraints, the GEM-PRO can

be used to simulate the effect of drug binding (Chang et al., 2013c, Aziz et al., 2015a, Aziz et al., 2015b), protein temperature sensitivity (Chang et al., 2013a), evolution of protein fold families (Zhang et al., 2009) and discovery of new metabolic activities (Guzman et al., 2015). Therefore GEM-PROs use the genotype to predict phenotypes of a cell that can then be validated experimentally (Monk and Palsson, 2014).

The functionality of GEM-PROs can be further extended by applying a range of in silico analytical methods (Lewis et al., 2012a). Recent studies have coupled the production of reactive oxygen species to the metabolic network, allowing predictions related to ROS production and detoxification as well as new ways to potentiate antibiotics (Brynildsen et al., 2013a). Also, GEM-PROs can serve as a scaffold for high-throughput data integration (Hyduke et al., 2013). This use allows interpretation of large datasets based on their effect on the network as a whole.

This study describes the first manually curated, genome-scale, elementally and charge balanced metabolic reconstruction combined with protein structures (GEM-PRO) for the important pathogen *S. aureus* USA300, termed *i*USA300_853-GP. This GEM-PRO allows for the formulation of hypothesis ranging from relative growth capabilities on different media to the outcome of gene deletion experiments and design of hypothetical new drugs. Importantly, due to the curation and refinement necessary to form a functional GEMPRO for *S. aureus* USA300, the work reported contains the most comprehensive metabolic reconstruction available for this significant pathogen that is consistent with known phenotypic functions.

8.3 Results

8.3.1 Reconstruction process and properties

The network reconstruction of *S. aureus* USA300 began with the metabolic network of the *S. aureus* N315 strain published in 2005 (Becker and Palsson, 2005). Further reconstruction data were obtained from two other metabolic networks (Heinemann et al., 2005, Lee et al., 2009) as well as literature and database searches. The MetaCyc (Keseler et al., 2009), KEGG (Kanehisa et al., 2010) and model SEED (Henry et al., 2010c) databases were used to examine newly characterized genes and reactions specific to the USA300 strain. All manual curation followed a standardized, well-established protocol (Thiele and Palsson, 2010b).

The reconstruction of cell wall and membrane metabolism was also substantially improved in the current reconstruction. The *S. aureus* cell wall is known to be important for bacterial resistance (Kuroda et al., 2003, Pechous et al., 2004) and cell wall structure and composition is known to fluctuate based on metabolic alterations (Cui et al., 2000). iJM755 has 242 reactions assigned to cell wall and membrane metabolism. These include strain specific reactions for the specialized enzymes required to synthesize the lipoteichoic and wall-teichoic acids, the composition of which are known to differ by genus (Neuhaus and Baddiley, 2003, Reusch, 1984, Sutcliffe and Shaw, 1991) in contrast to the pathways leading to synthesis of peptidoglycan precursors, which are well conserved between Gram-positive organisms (Schleifer and Kandler, 1972).

Detailed reconstruction of the respiratory chain was also performed as the iSB615 model was determined to incorrectly account for this feature of the organism. *S. aureus* is capable of growing aerobically and thus has a full respiratory chain including membrane-associated dehydrogenases, quinone electron transfer and ATPase. Menaquinone is known to be the only quinone in *S. aureus* (McNamara and Proctor,

2000, Tynecka et al., 1999, Wilkinson et al., 1997). Therefore, the ubiquinone dependency of the NADH dehydrogenase present in iSB615 was corrected.

The efficiency of ATPase in *S. aureus* (F0F1) was taken to be 2.4H+/ATP (2.4 protons transported across the membrane to synthesize one molecule of ATP), as determined by Heineman et al based on an experiment that determined the proton motive force in growing *Streptococcus* (Kashket, 1981) as well as applied experimental data on ATP, ADP and Pi concentrations for growing *S. aureus* cells (Tynecka et al., 1999, Christian and Waltho, 1964). This leads to approximately 1.7 molecules of ATP produced per mol of NADH, in agreement with the generally assumed ratio of 2 for *S. aureus (Wilkinson et al., 1997)*.

We also added the full staphyoxanthin pathway (Kim and Lee, 2012) to the model. Staphyloxanthin is a C(30) catotenoid biosynthesized by S. aureus. This pigment acts as an antioxidant and its numerous conjugated double bonds enabling the detoxification of host immune system–generated reactive oxygen species (ROS) such as O_2^-, H_2O_2, and HOCl (Clauditz et al., 2006, Liu et al., 2005).

A full annotation of known S. aureus virulence factors present in USA300 was also performed. Known staph aureus virulence factors were downloaded from databases (Chen et al., 2015) and identified from a literature search (Nizet, 2007). Overall we identified a total of 85 established virulence factors in USA300. We also identified genes involved in transcription and translation for this organism making future integration with an expanded model of metabolism and expression mechanisms (O'Brien et al., 2013b, Liu et al., 2014a) (ME-model) possible.

8.3.2 Integration with 3D protein structures

Next, this reconstruction of metabolic, virulence and expression capabilities of *S. aureus* USA300 was integrated with 3-dimensional protein structures. A standardized workflow was used to search the PDB for matching content. Missing content was added by identifying the nearest homolog with an existing structure. Homology models were built from this template modified to match the amino acid sequence of the USA300 query protein. Overall 159 proteins were found in the PDB and 790 proteins required homology modelling. Of those that require homology modeling, we have modeled 75% (589 non-transport related proteins).

The *i*USA300_853-GP reconstruction represents a significant expansion beyond any other *S. aureus* reconstruction as it contains 853 genes, 1,558 metabolic reactions, and 1,338 unique metabolites. The genome sequence of *S. aureus* USA300 (isolate FPR3757, NCBI: NC_007793) has a length of 2,872,769 base pairs with 2,560 predicted protein coding regions (Diep et al., 2006). Thus, our metabolic model covers approximately 33% of all ORFs. A comparison of the content of *i*USA300_853-GP and other *S. aureus* metabolic models is presented in **Supplementary Figure 1**. *i*USA300_853-GP covers a wide range of metabolic functions (**Figure 1**). The complete lists of reactions and metabolites in *i*USA300_853-GP can be found in **Supplementary Table 1**, with a list of all references used in **Supplementary Table 2**. The *i*USA300_853-GP model accounts for two compartments: the cytoplasm and extracellular space. Thus, *i*USA300_853-GP represents the most complete metabolic reconstruction of *S. aureus* USA300 available to date.

8.3.4 Biomass objective function

The biomass objective function (BOF) from *i*SB619 was updated in *i*USA300_853. The BOF is a reaction that drains biomass precursor compounds in

experimentally determined ratios to simulate growth (Feist and Palsson, 2010b). Both core and wild-type biomass functions were defined for S. aureus USA300. No data on biomass composition were available specifically for strain USA300. Hence, literature data from a variety of *S. aureus* strains were used to refine the biomass composition to form an average biomass function representative of the S. aureus species. The ratio of nucleic acids was determined by the G/C content of the genome sequence. Amino acid content was determined by assessing the average amino acid composition of genes in the genome as well as the pool of free solutes previously measured (Graham and Wilkinson, 1992, Hancock, 1960, Heinemann et al., 2005). The wild-type biomass function contains fatty acid composition of lipids required for growth specific to *S. aureus (Theodore and Panos, 1973)*. The most prominent lipids in *S. aureus* are phsopholipids (phosphatidyl glycerol, cardiolipin and lysyl-phosphatidyl glycerol (White and Frerman, 1967, Hugo and Davidson, 1973b, Koch et al., 1984), glycolipids (monoglycosyldiacyl glycerol, diglycosyldiacyl dlycerol and lipoteichoic acid (Hugo and Davidson, 1973a) and apolar lipids (mainly menaquninone (White and Frerman, 1967)) as well as 1,2-diacylglycerol (Koch et al., 1984).

Intracellular concentration of metal ions and ATP were also taken from existing literature (Christian and Waltho, 1964, Graham and Wilkinson, 1992, Vinnikov, 1988). Furthermore, new ATP maintenance parameters were defined. Not all consumed substrate is used for synthesis of new biomass and energy, therefore an estimation of “maintenance energy” (Varma et al., 1993) is required to account for ATP energy used to maintain the cell’s integrity. The ATP maintenance requirements are generally divided into growth associated (GAM) and non-growth associated (NGAM) maintenance. These parameters have not been determined for *S. aureus*, so we used experimentally

determined maintenance parameters determined from other organisms to assign these rates (NGAM: 5mmol ATP/(g cell dry weight), GAM: 40 mmol ATP/(g cell dry weight)/h). Our maintenance energy was dramatically different from those in iSB615 where GAM was set to 20,000 ATP/(g cell dry weight)/h.

8.3.5 Prediction of metabolic phenotypes

Flux balance analysis (FBA) (Orth et al., 2010b) can be used with a constraint-based model to predict metabolic flux distributions, growth rates, substrate uptake rates, and product secretion rates. The *i*JM755 model was used to make phenotypic predictions such as growth rates and central metabolic flux distributions. To demonstrate the utility of the *i*USA300_853-GP model in making these phenotypic predictions, we generated two large-scale sets of model phenotype predictions

8.3.6 Prediction of a defined minimal media for *S. aureus USA300*

The metabolic model was used to predict required growth supporting nutrients of S. aureus USA300. With glucose as a carbon source, ammonia as a nitrogen source, phosphourous and sulfate along with trace elements and minerals the metabolic model predicted that growth was not possible indicating S. aureus USA300 was missing biosynthetic pathways for components in its biomass function. Using this minimal media, production of each individual biomass component was simulated. It was found that the model could not produce thiamin due to a lack of thiamine-phosphate kinase (TMPK: 2.7.4.16), indicating that this strain is auxotrophic for thiamin. While our metabolic model predicts that USA300 can grow without addition of other amino acids, it is known that S. aureus exhibits complicated regulatory behavior related to amino acid growth. We demonstrated that addition of threonine, proline, serine, leucine and thiamin

to minimal M9-glucose media was sufficient to enable experimental growth of *S. aureus* USA300.

8.3.7 Prediction and validation of S. aureus USA300 catabolic capabilities for alternative nutrient sources

FBA was used to predict growth capabilities of *i*USA300_853 on alternative sole sources of carbon, nitrogen, phosphorus, and sulfur. The prediction were validated by comparison to a biolog phenotypic microarray experimental screen (**Figure 2A**). In total the model was predicted to be capable of catabolizing 113 different nutrient sources 67 of which (65%) were experimentally confirmed. Those nutrient sources that did not show experimental growth may have be disabled due to regulatory restrictions that are outside the scope of this metabolic model (Orth and Palsson, 2010a, O'Brien et al., 2015). The model incorrectly predicted that S. aureus could not catabolize 28 different nutrients. These cases represent an opportunity for future model improvement. The model correctly predicted that *S. aureus* could not catabolize 86 nutrient sources (75%). The overall prediction accuracy for ability to use different nutrient sources was 73% which is in line with metabolic models of other organisms (Monk and Palsson, 2014).

8.3.8 Prediction and validation of essential genes in S. aureus USA300

An *in-silico* screen of model predicted growth capabilities for all possible single and double gene knockout strains was performed. Growth phenotypes were predicted on defined rich media, and the results for single gene knockouts were compared with with an experimental transposon mutagenesis dataset (Chaudhuri et al., 2009). A total of 83 metabolic genes were predicted to be essential, 65 of which (78%) agreed with experimental data. The incorrect predictions indicate missing model content and could be explained by alternate pathways or isozymes not currently represented in the model

and thus an opportunity for model expansion and improvement. Also, such false negative predictions have proven to be a way to target and discover new metabolic functions (Orth and Palsson, 2012, Guzman et al., 2015). The model predicted that 665 genes were nonessential for growth, 562 of which (85%) aligned with experimental data. The false positive predictions (model predicts growth but no growth is observed) are likely the results of regulatory restrictions or other factors that are outside the scope of the model. It is possible that continued culture of these gene knockout strains would grow eventually (Lee and Palsson, 2010). Overall the model correctly predicted the lethality of 627 (83%) of gene knockouts. This accuracy is in line with other metabolic models.

8.3.9 Integration of GEM-PRO with high-throughput data sets elucidates antibiotic mechanisms

Beyond acting as a predictive model of an organism's metabolic capabilities, a GEM-PRO can also serve as a scaffold for integration of high-throughput data. Such integration allows for a deeper interpretation of the observed changes on a systems level. To illustrate this capability, the RNA-seq omics transcriptome profiling technique was used to measure genome-wide transcriptomic profiles of *S. aureus* USA300 after treatment with two different drugs, Naficillin and the immune defense factor human cathelicidin LL-37. We compared these transcription profiles to wild-type transcription data. Figure 3 shows the gene expression comparison between wild type and treatment with either LL-37 (Figure 3A) or Naficillin (Figure 3B). We classified the genes into categories of metabolic (M, blue), expression (E, green) and virulence (V, red).

Under treatment with immune defense factor LL-37 there were 152 genes significantly differentially expressed ($p<0.05$) compared to wildtype conditions

(Supplementary Datafile X). Of these 152 genes, 43 were metabolic genes (21 up-regulated, 22 down regulated) including the respiratory nitrate and nitrite reductase genes *nar*, *nir* (all log2 FC> 4.5). There were 42 E genes differentially expressed (41 up-regulated, 1 down-regulated) including mostly 30S and 50S ribosomal proteins. Finally, only 5 virulence genes were differentially expressed. Most of these (4) were up-regulated including fibrinogen binding protein, *efb* (log2 FC: 3.8), fibronectin binding protein B, *fnbB* (log2 FC: 2.0), staphylocoagulase, *coa* (log2 FC: 3.8), and *sdrD (log2 FC: 3.1) (*read victor's review on this). Meanwhile, delta-hemolysin was down-regulated (log2 FC: -3.1). The 62 genes that were not metabolic, expression or virulence related were primarily hypothetical proteins (37%).

Treatment with Naficillin showed 230 genes significantly differentially expressed ($p<0.05$) compared to wildtype conditions. Of these 230 genes, 64 were metabolic genes (15 up-regulated, 49 down regulated) again including the respiratory nitrate and nitrite reductase genes *nar*, *nir* (all log2 FC> 4.5). There were 25 E genes differentially expressed (19 up-regulated, 6 down-regulated) including mostly 30S and 50S ribosomal proteins. Finally, only 7 virulence genes were differentially expressed. Most of these (5) were up-regulated including penicillin binding protein *mecA* (Log2 FC: 5.4) , fibrinogen binding protein, *efb* (log2 FC: 2.7), staphylocoagulase, *coa* (log2 FC: 3.0), and the adherence genes *sdrC (log2 FC: 2.3) and sdrD (log2 FC: 3.9)*. Meanwhile, delta-hemolysin (log2 FC: -7.4) and the type VII secretion system gene *esxA* (log2 FC: -2.2) were downregulated. The 134 genes that were not metabolic, expression or virulence related were primarily hypothetical proteins (43%), or involved in regulatory and sensory processes.

8.3.10 GEM-PRO enables prediction of reactive oxygen species production and detoxification in *S. aureus*

One application that is enabled with a full GEM-PRO for an organism is the the modelling and prediction of reactive oxygen species (ROS) production and detoxification. This approach has been applied and validated for the study of ROS production in *E. coli* (Brynildsen et al., 2013a). Metabolic reactions in the GEM can be coupled with known ROS producing sources. A major source of ROS in an organism are enzymes that use flavins, quninones, hemes and transition metals (Massey, 1994, Messner and Imlay, 1999). Three-dimensional representations of proteins conserved domain structures can be used to predict flavin, quinone, transition metal and heme binding sites on proteins. Thus, they can be used to predict enzymes that utilize these cofactors and may serve as endogenous sources of ROS production. We used the GEM-PRO of *S. aureus* to predict the enzymes in our metabolic network that bind bind flavins (34), quinones (7), hemes (19) and transition metals. We based these predictions on both protein sequence alignments with known protein fold families (PFAM), as well as secondary structural alignments, using the FATCAT algorithm (Ye and Godzik, 2003).

We used an established approach to couple the reactions catalyzed by these ROS producing enzymes with production of ROS species H2O2 and O2s (**Methods**). Thus, whenever flux flows through these reactions, the ROS species are produced as by-products as well. In total 122 reactions were coupled with ROS production. The rate at which each enzyme produces ROS is unknown, therefore we used an ensemble approach to generate 1000 different ROS production coefficients for each of the 122 reactions. The 1000 values were generated using both a gaussian distribution for an assumption of uniform ROS production across the reactions and an exponential distribution to model individual high ROS producers with a majority of lower producers.

A representative ROS-coupled metabolic model of *S. aureus* USA300 termed *i*USA300_853-ROS is available in **Supplementary File 2.**

Using this ensemble of ROS-coupled metabolic models we ran simulations of single gene knockouts one-by-one to identify those that reliably led to increases in ROS production. Such situations occur when a gene knockout leads to re-routing of metabolic flux through reactions that produce ROS, leading to an increase in overall ROS production by the network. Overall we found 24 high-confidence gene knockouts that led to 5% ROS production increase in over 70% of the ensemble models. 15 of these knockouts are shared with predictions for *E. coli*, 6 of which were experimentally shown to increase ROS production and potentiate antibacterial activity.

8.4 Discussion

S. aureus USA300 is a strain of community-associated MRSA that is the cause of a new antibiotic resistant epidemic responsible for rapidly progressive, fatal diseases. Starting in the late 1990s, the USA300 lineage of methicillin-resistant Staphylococcus aureus (MRSA) underwent an extremely rapid expansion across the United States, replacing many other S. aureus strains (Tenover and Goering, 2009). Since that time, it has become a major cause of skin and soft-tissue infections (Moran et al., 2006), community-acquired pneumonia, catheter-related bloodstream infections (Moran et al., 2006), and other systemic infections (Klevens et al., 2007). Outbreaks of community-associated (CA)-MRSA infections have been reported in correctional facilities, among athletic teams, among military recruits, in newborn nurseries, and among sexually active homosexual men. CA-MRSA infections now appear to be endemic in many urban regions and cause most MRSA infections in the United States (Maree et al., 2007, Diep et al., 2008).

Here we present a genome-scale metabolic network reconstruction of S. aureus USA300 integrated with three-dimensional protein structures of enzymes, virulence factors and proteins involved in transcription and translation. The network reconstruction covers the function of 853 genes. It can be converted to a genome-scale model using a mathematical representation that allows simulation of S. aureus USA300 metabolic capabilities and weaknesses based on its genotype. The model is capable of reproducing known S. aureus physiology including anaerobic fermentation of lactate and metabolism of small colon variants. We use the genome scale model to predict nutrients that *S. aureus* can use as sole sources of carbon, nitrogen, phosphorous and sulfur with 73% accuracy. Each of these capabilities is explicitly linked to its catabolic pathway, catalyzing reactions and encoding genes. This connection could allow for the design of bacteriostatic therapies that target an organism's ability to survive and thrive in its preferred infectious niche. We also use the model to predict essential single-gene and double-gene knockouts in rich media conditions. The single gene knockouts are predicted with an accuracy of 83% indicating that both the single gene predictions and double gene predictions are accurate enough to be used as targets for the design of anti-metabolite and anti-enzymatic inhibitors as well as synthetic lethal inhibitory combinations.

We further illustrate the utility of iUSA300_853 to act as a scaffold for high-throughput data integration. We transcriptionally profiled USA300 under treatment with two different antibiotics Nafcillin and LL-37 and overlaid the transcriptional changes onto the model compared to wild-type growing cells. We further integrated the model with ROS production reactions to predict the major source of ROS production and detoxification in S. aureus USA300. We used the ROS-coupled model to predict gene

knockouts that would reliably lead to increases in ROS production. Thus these predictions should be tested in S. aureus USA300 to may be used in conjunction with antibiotics to potentiate their activity and enhance killing effectiveness.

The GEM-PRO for S. aureus USA300 presented here is the most accurate and complete genome-scale reconstruction for any strain of S. aureus. It can be used for prediction of metabolic capabilities, essential genes, as a scaffold for high-throughput omics data integration, for the simulation of ROS production and detoxification and for the interpretation of genetic changes observed in clinical isolates of this rapidly spreading and constantly evolving pathogen. Its wide application and use should lead to a systems-level understanding of metabolism and virulence and will hopefully lead to new antibiotics and treatment therapies.

8.5 Acknowledgements

Chapter 8, in part, is a reprint of the material Monk JM, Mih N, Brunk E, Aziz RK, Palsson BO. A genome-scale reconstruction of metabolism and associated protein structures in *Staphlyococcus aureus* USA300. *In Preparation.* The disseration author was the primary author of this paper.

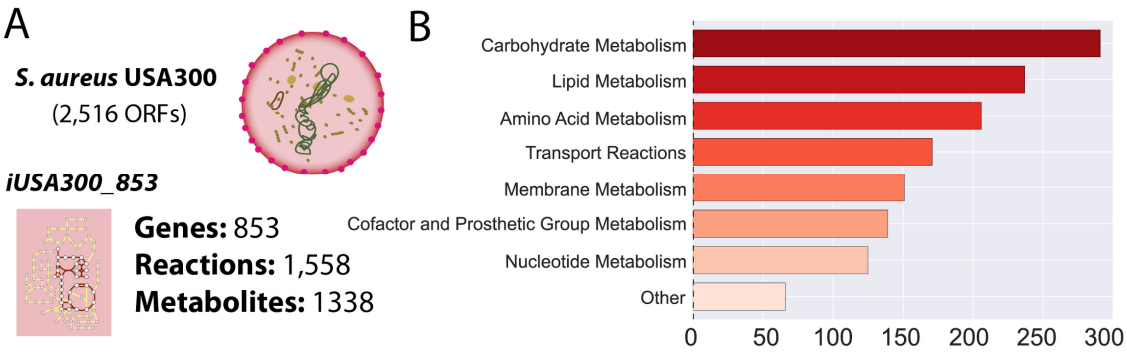


Figure 8.1. Properties of iUSA300_853-GP.
A) The *i*USA300_853-GP metabolic reconstruction covers the functions of 853 genes, 1,588 reactions and 1,338 metabolites. B) The reconstruction content is subdivided into different metabolic subsystems.

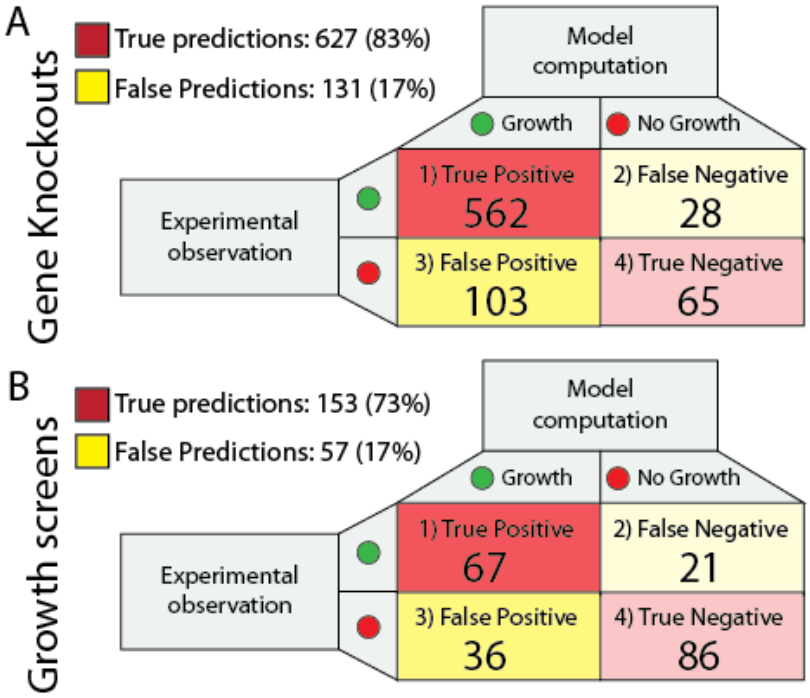


Figure 8.2. Experimental validation of model predictions. A) Gene knockout predictions. The model was used to predict the effect of each gene knockout performed individually when grown in rich media conditions. The results of this screen were compared to experimental genome-wide screens of gene essentiality. Overall the model correctly predicted 562 genes to be non-essential (model growth/experimental growth) and 65 genes to be essential (model no growth/experimental no growth). This represents an accuracy of 83%. B) Nutrient usage predictions. The model was used to predict growth capabilities on 210 different carbon, nitrogen, sulfur and phosphorous sources. Overall the model correctly predicted growth capabilities (growth/no growth) on 73% of the compounds.

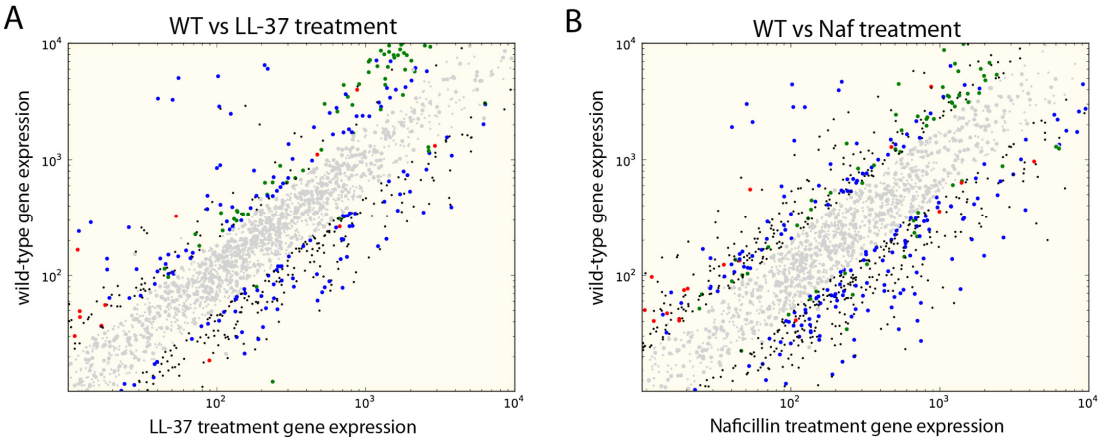


Figure 8.3. Gene expression profile comparison for S. aureus USA300 treated with two antibiotics:
The human immune defense factor human cathelicidin LL-37 and B) the beta lactam Naficillin. Both treatments are compared to wild-type expression profiles on the x-axis. Genes in blue are those in the metabolic model (M), green are genes involved in protein expression (transcription and translation) (E) and red are genes marked as virulence factors (V). Those in grey are not significantly differentially expressed (abs(log2 FC) < 1).

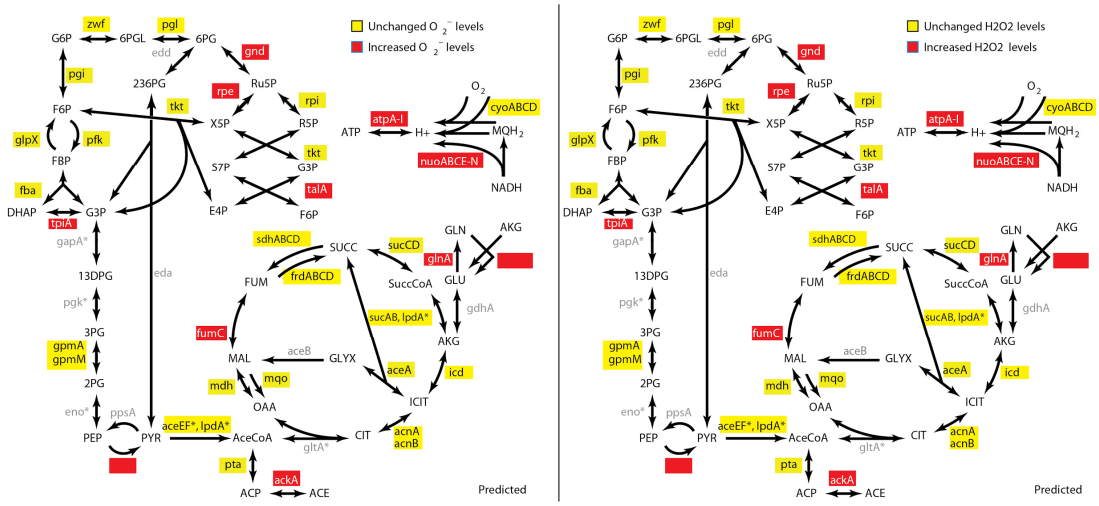


Figure 8.4. Model predicted gene knockouts in central metabolism that lead to increases in ROS production.

A) Model predicted gene knockouts that lead to an increase in O2- levels. **B)** Model predicted gene knockouts that lead to increases in H2O2 levels. Red indicates an increase of at least across more than of ensemble models. Yellow indicates no that ROS production was not predicted to increase or that any increase in production did not meet the previous criteria.

Chapter 9 Final Thoughts and Future Direction

9.1 Massive-scale assessment of phenotypic predictions will have broad consequences

The genotype-phenotype relationship is fundamental to biology. A cellular phenotype results from biochemical interactions amongst thousands of cellular components occurring in the confines of the cell. With the availability of annotated genome sequences and a plethora of new omics data types, we are now able to assemble these biochemical interactions on a genome-scale for model microorganisms. Such an assembly of detailed biochemical information can be converted into a computational model using a simple mathematical representation (no theory involved!) that in turn can be used to compute phenotypic functions. Both environmental and genetic parameters are explicitly accounted for in such *in silico* cellular models that enable predictions of the genotype-phenotype relationship in a given environment. Fifteen years ago, this pursuit may have sounded like a pipedream, but through the hard work of a small but growing community of researchers, such predictions have been realized.

Shortly after the sequencing of the first genomes, genome-scale models (GEMs) of metabolic functions appeared[1]. The basis for a GEM is a highly curated reconstruction of the underlying biochemical reaction networks that the organism's functions are based on. One can think of the process of network reconstruction in analogy to sequence assembly. The genome of a cell is assembled from many short DNA stubs, called reads, using sophisticated computer algorithms. Similarly, the reactome of a cell is assembled, or reconstructed, from all the biochemical reactions known or predicted to be present in the target microorganism. Importantly, the network reconstruction includes the genetic basis for each reaction in the reactome.

The inner workings of a GEM are readily understood on a conceptual basis. A GEM is a mathematical representation of a reconstructed network. In a given environment (i.e., where the nutritional inputs are defined) GEMs can be used to compute network outputs (a phenotype), such as the formation of biomass. The synthesis of biomass in a cell requires about 60-70 different metabolites[2]. A GEM can compute a path to the biosynthesis of each one of these prerequisite metabolites by computationally tracing a fully balanced path through the reactome from the available nutrients to the prerequisite metabolite. A GEM can also compute the balanced use of the reactome to form all the prerequisite metabolites simultaneously and in the correct relative amounts while accounting for all the energetic, redox and chemical interactions that must balance to enable such biomass synthesis. Such is the power of bottom-up systems biology. It may seem like magic, but conceptually speaking, this is simply a genome-scale accounting exercise that is predicated on the reconstruction of an accurate reactome. The challenges of building accurate reactomes have been recently discussed and the range of phenotypes that can be computed has been described[3,4]. Here we will focus on the prediction of possible cellular growth states and the implications of scaling up the number of such predictions. Similar considerations apply to other measureable phenotypes that are computable.

Underlying the computation of a growth state is a simple model of inputs and outputs. The inputs pass through a GEM's internal reactome that is comprised of reactions linked to genes, giving them an explicit genetic basis (**figure panel A**). The simplicity of computing growth states (i.e., an output) as a function of media composition (i.e., the nutritional inputs) with the selective removal of genes has led to a number of studies that cross environmental parameters (E) with gene deletions (G). This explicit

relationship between a gene and a reaction makes the deletion of genes and their encoding reactions straightforward.

Phenotypic ExG growth screens have been performed with target organisms for which gene knock-out (KO) strain collections exist, such as *Escherichia coli*[5], *Saccharomyces cerevisiae*[6] and *Bacillus subtilis*[7]. GEMs can compute the outputs of such experiments. The number of phenotypic predictions with experimental validation from ExG screens has grown steadily over the past 15 years (**figure panel B**). The initial comparisons included less than 100 predictions, and the more recent ones exceed 100,000 predictions. *This series of studies represents the largest-scale attempts at bona fide predictions of phenotypic outcomes by a computational model.*

Computational predictions of outcomes fall into four categories. The true-positive and true-negative predictions generally exceed 80% to 90% for the organisms examined. For double knockouts, true negative predictions are particularly significant as they indicate model predictions of true genetic interactions. Highly curated models can predict up to 50% of such interactions and the missed predictions represent cases that are currently difficult for functional genomicists to understand[8]. For this reason, the failure of prediction is perhaps of more interest as it represents an opportunity for true biological discovery. False negative predictions occur when a GEM predicts the inability to grow in an environment tested without the deleted gene, but the experiment results in growth. This discrepancy indicates that the reconstructed reactome is incomplete. In contrast, false positive predictions occur when a GEM predicts growth but the experiment results in no growth. This difference indicates possible errors in the knowledge that the reactome was based on, or that a regulatory process is missing; for example, regulation that either represses gene expression or a metabolite-enzyme

interaction that inhibits the function of an enzyme that the GEM used to compute the predicted growth state.

Reconciling such discrepancies between predicted and observed growth states is now a proven approach for biological discovery. A series of algorithms have been developed that have been shown to compute the most likely reasons for failure of prediction that in turn led to a model-guided experimental inquiry and discovery[9]. A few recent studies illustrate how this approach works. Two new reactions carried out by the classical enzymes phosphofructokinase and aldolase were discovered through such systematic inquiry in *E. coli*[10]. Corrections of the pathways leading to NAD synthesis in yeast also resulted from analysis of synthetic lethal screens in yeast[8] and gluconate kinase was described in human[11] (**figure panel C**). Fortunately, failure modes are not independent and many false predictions can be resolved by fixing relatively few components in a model's reactome. Recently a reconciliation of 2,442 false model predictions was obtained for the *E. coli* GEM by updating the function of just 12 genes[12].

Thus, a GEM provides a platform not only for formalizing and solidifying our understanding of a target organism, but also for the systematic discovery of its missing parts and functions. With double KO collections being produced for *E. coli* where highly quantitative measurements under defined growth conditions can be obtained, we can foresee millions of growth predictions being possible in just a few years. Such experiments and computational predictions would enable a large-scale prediction of epistatic interactions and would represent an unprecedented probing of our understanding of the genotype-phenotype relationship for a target organism.

If such efforts for model organisms are successful, and if genome-editing tools for lesser-characterized organisms become readily available, then we can foresee a

massive scale-up in phenotypic predictions for a spectrum of organisms. For instance, transposon libraries with sequencing (TnSeq) can now be used to examine the effect of gene knockouts in different environments for organisms with little to no existing legacy biochemical data[13]. A draft GEM for such an organism can thus be improved quickly using the same computational tools mentioned above. Such developments would hopefully make microbial physiologists just as happy as microbial genomicists have become over the past 15 years. If target organisms can be strategically chosen from across the phylogenetic tree, we should be able to comprehensively discover the metabolic processes resident on earth that support growth of diverse organisms, and similarly the genotypic basis for other assayable phenotypic functions[3,4]. If pursued and accomplished, such an undertaking would represent the resolution of a grand challenge in biology.

9.2 Acknowledgements

Chapter 9, in part, is a reprint of the material Monk JM, Palsson BO: Genetics. Predicting microbial growth. *Science* 2014, 344(6191):1448-1449.The disseration author was the primary author of this paper.

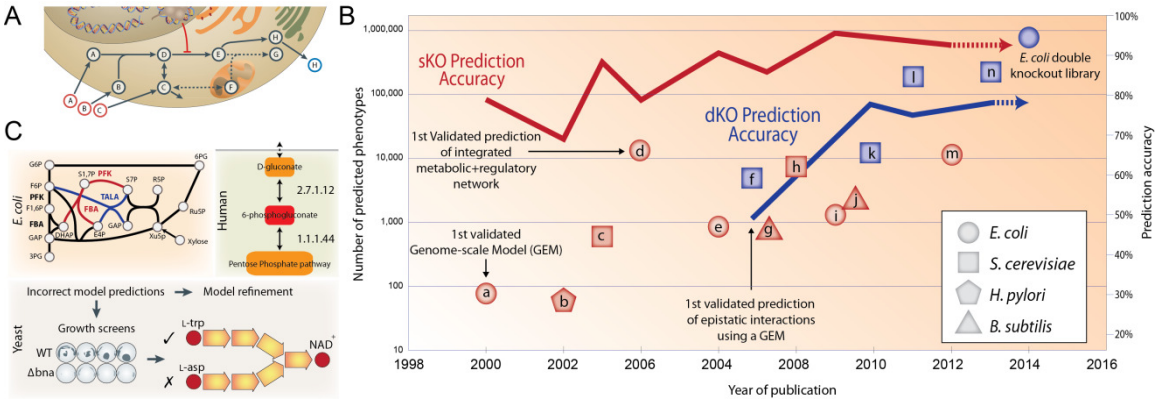


Figure 9.1. Applications of genome scale modelling and its increases in predictive accuracy over time

A) Genome-scale models (GEMs) are Input-Output (I/O) models where genetic knockouts (red) change internal reaction structure of the system. **B)** Historical profile of studies that use GEMs to predict experimental outcomes of phenotypic screens. The number of phenotypic predictions from ExG screens has grown steadily over the past 15 years. The red line indicates accuracy of single gene knockout (SKO) prediction accuracy and the blue line indicates prediction of double gene knockout (DKO) accuracy. **C)** The comparison of GEM computation and organism-specific experimental measurements identifies agreements and disagreements. The resolution of these disagreements can lead to new biological discovery, clockwise from upper left are examples from three model organisms: *E. coli*, human and yeast. *E. coli*: Two new functions for the classical biochemistry enzymes phosphofructokinase (PFK) and fructose-bisphosphate aldolase (FBA) were discovered (red)[10]. Also the *E. coli* talA and talB genes were newly discovered to catalyze a transaldolase reaction TALA (blue) based on knockout phenotypes. Human: Gluconokinase (EC 2.7.1.12) activity was discovered based on the known presence of the metabolite 6-phosphogluconolactonate in the human reconstruction[11] (red). Yeast: Automated model refinement suggested modifications in NAD biosynthesis pathway. Experimental results indicated negative genetic interactions in the NAD biosynthesis pathway (Δbna vs WT) starting from tryptophan indicating that a parallel pathway from aspartate thought to exist in yeast was not present.

Bibliography

ADKINS, J., PUGH, S., MCKENNA, R. & NIELSEN, D. R. 2012. Engineering microbial chemical factories to produce renewable "biomonomers". *Front Microbiol,* 3**,** 313.

ADOLFSEN, K. J. & BRYNILDSEN, M. P. 2015. Futile cycling increases sensitivity toward oxidative stress in Escherichia coli. *Metab Eng,* 29**,** 26-35.

AGARWAL, R., BURLEY, S. K. & SWAMINATHAN, S. 2007. Structural analysis of a ternary complex of allantoate amidohydrolase from Escherichia coli reveals its mechanics. *J Mol Biol,* 368**,** 450-63.

AGREN, R., BORDEL, S., MARDINOGLU, A., PORNPUTTAPONG, N., NOOKAEW, I. & NIELSEN, J. 2012. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology,* 8**,** e1002518.

AGREN, R., LIU, L., SHOAIE, S., VONGSANGNAK, W., NOOKAEW, I. & NIELSEN, J. 2013a. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS computational biology,* 9**,** e1002980.

AGREN, R., LIU, L., SHOAIE, S., VONGSANGNAK, W., NOOKAEW, I. & NIELSEN, J. 2013b. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLoS Comput Biol,* 9**,** e1002980.

AKANUMA, G., NANAMIYA, H., NATORI, Y., YANO, K., SUZUKI, S., OMATA, S., ISHIZUKA, M., SEKINE, Y. & KAWAMURA, F. 2012. Inactivation of ribosomal protein genes in Bacillus subtilis reveals importance of each ribosomal protein for cell proliferation and cell differentiation. *J Bacteriol,* 194**,** 6282-91.

ARCHER, C. T., KIM, J. F., JEONG, H., PARK, J. H., VICKERS, C. E., LEE, S. Y. & NIELSEN, L. K. 2011. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics,* 12**,** 9.

ARIFIN, Y., ARCHER, C., LIM, S., QUEK, L. E., SUGIARTO, H., MARCELLIN, E., VICKERS, C. E., KROMER, J. O. & NIELSEN, L. K. 2014. Escherichia coli W shows fast, highly oxidative sucrose metabolism and low acetate formation. *Appl Microbiol Biotechnol,* 98**,** 9033-44.

ATLAS, R. 2010. *Handbook of Microbiological Media*, CRC Press.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A. & ZAGNITKO, O.

2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* 9**,** 75.

AZIZ, R. K., DEVOID, S., DISZ, T., EDWARDS, R. A., HENRY, C. S., OLSEN, G. J., OLSON, R., OVERBEEK, R., PARRELLO, B., PUSCH, G. D., STEVENS, R. L., VONSTEIN, V. & XIA, F. 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One,* 7**,** e48053.

AZIZ, R. K., KHAW, V. L., MONK, J. M., BRUNK, E., LEWIS, R., LOH, S. I., MISHRA, A., NAGLE, A. A., SATYANARAYANA, C., DHAKSHINAMOORTHY, S., LUCHE, M., KITCHEN, D. B., ANDREWS, K. A., PALSSON, B. O. & CHARUSANTI, P. 2015a. Model-driven discovery of synergistic inhibitors against E. coli and S. enterica serovar Typhimurium targeting a novel synthetic lethal pair, aldA and prpC. *Front Microbiol,* 6**,** 958.

AZIZ, R. K., MONK, J. M., LEWIS, R. M., IN LOH, S., MISHRA, A., ABHAY NAGLE, A., SATYANARAYANA, C., DHAKSHINAMOORTHY, S., LUCHE, M., KITCHEN, D. B., ANDREWS, K. A., FONG, N. L., LI, H. J., PALSSON, B. O. & CHARUSANTI, P. 2015b. Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. *Sci Rep,* 5**,** 16025.

BABA, T., ARA, T., HASEGAWA, M., TAKAI, Y., OKUMURA, Y., BABA, M., DATSENKO, K. A., TOMITA, M., WANNER, B. L. & MORI, H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol,* 2**,** 2006 0008.

BANCROFT, E. A. 2007. Antimicrobial resistance: it's not just for hospitals. *JAMA,* 298**,** 1803-4.

BARAN, R., BOWEN, B. P., PRICE, M. N., ARKIN, A. P., DEUTSCHBAUER, A. M. & NORTHEN, T. R. 2013. Metabolic footprinting of mutant libraries to map metabolite utilization to genotype. *ACS Chem Biol,* 8**,** 189-99.

BARUA, D., KIM, J. & REED, J. L. 2010. An automated phenotype-driven approach (GeneForce) for refining metabolic and regulatory models. *PLoS Comput Biol,* 6**,** e1000970.

BASEMAN, J. B. & COX, C. D. 1969. Intermediate energy metabolism of Leptospira. *J Bacteriol,* 97**,** 992-1000.

BAUMLER, D. J., PEPLINSKI, R. G., REED, J. L., GLASNER, J. D. & PERNA, N. T. 2011. The evolution of metabolic networks of E. coli. *BMC Syst Biol,* 5**,** 182.

BECKER, S. A. & PALSSON, B. O. 2005. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol,* 5**,** 8.

BECKER, S. A. & PALSSON, B. O. 2008. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology,* 4**,** e1000082.

BEG, Q. K., VAZQUEZ, A., ERNST, J., DE MENEZES, M. A., BAR-JOSEPH, Z., BARABASI, A. L. & OLTVAI, Z. N. 2007. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 12663-8.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2009. GenBank. *Nucleic Acids Res,* 37**,** D26-31.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & WHEELER, D. L. 2005. GenBank. *Nucleic Acids Res,* 33**,** D34-8.

BENTLEY, R. & MEGANATHAN, R. 1982. Biosynthesis of vitamin K (menaquinone) in bacteria. *Microbiol Rev,* 46**,** 241-80.

BERNIER-FEBREAU, C., DU MERLE, L., TURLIN, E., LABAS, V., ORDONEZ, J., GILLES, A. M. & LE BOUGUENEC, C. 2004. Use of deoxyribose by intestinal and extraintestinal pathogenic Escherichia coli strains: a metabolic adaptation involved in competitiveness. *Infect Immun,* 72**,** 6151-6.

BHARTI, A. R., NALLY, J. E., RICALDI, J. N., MATTHIAS, M. A., DIAZ, M. M., LOVETT, M. A., LEVETT, P. N., GILMAN, R. H., WILLIG, M. R., GOTUZZO, E. & VINETZ, J. M. 2003. Leptospirosis: a zoonotic disease of global importance. *Lancet Infect Dis,* 3**,** 757-71.

BINTER, E., BINTER, S., DISZ, T., KALMANEK, E., POWERS, A., PUSCH, G. & TURGEON, J. 2012. Grounding annotations in published literature with an emphasis on the functional roles used in metabolic models. *3 Biotech.,* 2**,** 135-140.

BLATTNER, F. R., PLUNKETT, G., 3RD, BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. & SHAO, Y. 1997. The complete genome sequence of Escherichia coli K-12. *Science,* 277**,** 1453-62.

BLIVEN, K. A. & MAURELLI, A. T. 2012. Antivirulence genes: insights into pathogen evolution through gene loss. *Infect Immun,* 80**,** 4061-70.

BOHMER, A., MULLER, A., PASSARGE, M., LIEBS, P., HONECK, H. & MULLER, H. G. 1989. A novel L-glutamate oxidase from Streptomyces endus. Purification and properties. *Eur J Biochem,* 182**,** 327-32.

BORDBAR, A., LEWIS, N. E., SCHELLENBERGER, J., PALSSON, B. O. & JAMSHIDI, N. 2010. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Molecular systems biology,* 6**,** 422.

BORDBAR, A., MONK, J. M., KING, Z. A. & PALSSON, B. O. 2014. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet,* 15**,** 107-20.

BORDBAR, A. & PALSSON, B. O. 2012. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *J Intern Med,* 271**,** 131-141.

BRAUDE, A. I. & SIEMIENSKI, J. 1960. Role of bacterial urease in experimental pyelonephritis. *J Bacteriol,* 80**,** 171-9.

BRYNILDSEN, M. P., WINKLER, J. A., SPINA, C. S., MACDONALD, I. C. & COLLINS, J. J. 2013a. Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. *Nat Biotechnol,* 31**,** 160-5.

BRYNILDSEN, M. P., WINKLER, J. A., SPINA, C. S., MACDONALD, I. C. & COLLINS, J. J. 2013b. Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. *Nature biotechnology,* 31**,** 160-5.

BUBUNENKO, M., BAKER, T. & COURT, D. L. 2007. Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in Escherichia coli. *J Bacteriol,* 189**,** 2844-53.

BURGARD, A. P., PHARKYA, P. & MARANAS, C. D. 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering,* 84**,** 647-57.

BUYUKTIMKIN, B. & SAIER, M. H., JR. 2015. Comparative genomic analyses of transport proteins encoded within the genomes of Leptospira species. *Microb Pathog,* 88**,** 52-64.

BYRGAZOV, K., VESPER, O. & MOLL, I. 2013. Ribosome heterogeneity: another level of complexity in bacterial translation regulation. *Curr Opin Microbiol,* 16**,** 133-9.

CAIMANO, M. J., SIVASANKARAN, S. K., ALLARD, A., HURLEY, D., HOKAMP, K., GRASSMANN, A. A., HINTON, J. C. & NALLY, J. E. 2014. A model system for studying the transcriptomic and physiological changes associated with mammalian host-adaptation by Leptospira interrogans serovar Copenhageni. *PLoS Pathog,* 10**,** e1004004.

CAMPODONICO, M. A., ANDREWS, B. A., ASENJO, J. A., PALSSON, B. O. & FEIST, A. M. 2014. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. *Metab Eng,* 25**,** 140-58.

CANCHAYA, C., PROUX, C., FOURNOUS, G., BRUTTIN, A. & BRUSSOW, H. 2003. Prophage genomics. *Microbiol Mol Biol Rev,* 67**,** 238-76, table of contents.

CARTER, E. L., JAGER, L., GARDNER, L., HALL, C. C., WILLIS, S. & GREEN, J. M. 2007. Escherichia coli abg genes enable uptake and cleavage of the folate catabolite p-aminobenzoyl-glutamate. *J Bacteriol,* 189**,** 3329-34.

CASPI, R., ALTMAN, T., BILLINGTON, R., DREHER, K., FOERSTER, H., FULCHER, C. A., HOLLAND, T. A., KESELER, I. M., KOTHARI, A., KUBO, A., KRUMMENACKER, M., LATENDRESSE, M., MUELLER, L. A., ONG, Q., PALEY, S., SUBHRAVETI, P., WEAVER, D. S., WEERASINGHE, D., ZHANG, P. & KARP, P. D. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res,* 42**,** D459-71.

CASPI, R., FOERSTER, H., FULCHER, C. A., KAIPA, P., KRUMMENACKER, M., LATENDRESSE, M., PALEY, S., RHEE, S. Y., SHEARER, A. G., TISSIER, C., WALK, T. C., ZHANG, P. & KARP, P. D. 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res,* 36**,** D623-31.

CHAE, H. S., KIM, K. H., KIM, S. C. & LEE, P. C. 2010. Strain-dependent carotenoid productions in metabolically engineered Escherichia coli. *Appl Biochem Biotechnol,* 162**,** 2333-44.

CHAMBERS, H. F. & DELEO, F. R. 2009. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nat Rev Microbiol,* 7**,** 629-41.

CHANDRASEKARAN, S. & PRICE, N. D. 2010. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 17845-50.

CHANG, R. L., ANDREWS, K., KIM, D., LI, Z., GODZIK, A. & PALSSON, B. O. 2013a. Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. *Science,* 340**,** 1220-3.

CHANG, R. L., GHAMSARI, L., MANICHAIKUL, A., HOM, E. F. Y., BALAJI, S., FU, W., SHEN, Y., HAO, T., PALSSON, B. O., SALEHI-ASHTIANI, K. & PAPIN, J. A. 2011. Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol Syst Biol,* 7.

CHANG, R. L., XIE, L., BOURNE, P. E. & PALSSON, B. O. 2010. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS computational biology,* 6**,** e1000938.

CHANG, R. L., XIE, L., BOURNE, P. E. & PALSSON, B. O. 2013b. Antibacterial mechanisms identified through structural systems pharmacology. *BMC systems biology,* 7**,** 102.

CHANG, R. L., XIE, L., BOURNE, P. E. & PALSSON, B. O. 2013c. Antibacterial mechanisms identified through structural systems pharmacology. *BMC Syst Biol,* 7**,** 102.

CHARUSANTI, P., CHAUHAN, S., MCATEER, K., LERMAN, J. A., HYDUKE, D. R., MOTIN, V. L., ANSONG, C., ADKINS, J. N. & PALSSON, B. O. 2011. An experimentally-supported genome-scale metabolic network reconstruction for Yersinia pestis CO92. *BMC Syst Biol,* 5**,** 163.

CHAUDHURI, R. R., ALLEN, A. G., OWEN, P. J., SHALOM, G., STONE, K., HARRISON, M., BURGIS, T. A., LOCKYER, M., GARCIA-LARA, J., FOSTER, S. J., PLEASANCE, S. J., PETERS, S. E., MASKELL, D. J. & CHARLES, I. G. 2009. Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics,* 10**,** 291.

CHEN, J., BERGEVIN, J., KISS, R., WALKER, G., BATTISTONI, T., LUFBURROW, P., LAM, H. & VINTHER, A. 2012a. Case Study: A Novel Bacterial Contamination in Cell Culture Production--Leptospira licerasiae. *PDA J Pharm Sci Technol,* 66**,** 580-91.

CHEN, L. & VITKUP, D. 2007. Distribution of orphan metabolic activities. *Trends Biotechnol,* 25**,** 343-348.

CHEN, L., XIONG, Z., SUN, L., YANG, J. & JIN, Q. 2012b. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res,* 40**,** D641-5.

CHEN, L., ZHENG, D., LIU, B., YANG, J. & JIN, Q. 2015. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Res*.

CHRISTIAN, J. H. & WALTHO, J. A. 1964. The Composition of Staphylococcus Aureus in Relation to the Water Activity of the Growth Medium. *J Gen Microbiol,* 35**,** 205-13.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B. & BORK, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science,* 311**,** 1283-7.

CLARKE, S. R. & FOSTER, S. J. 2008. IsdA protects Staphylococcus aureus against the bactericidal protease activity of apolactoferrin. *Infect Immun,* 76**,** 1518-26.

CLAUDITZ, A., RESCH, A., WIELAND, K. P., PESCHEL, A. & GOTZ, F. 2006. Staphyloxanthin plays a role in the fitness of Staphylococcus aureus and its ability to cope with oxidative stress. *Infect Immun,* 74**,** 4950-3.

COLLAKOVA, E., YEN, J. Y. & SENGER, R. S. 2012. Are we ready for genome-scale modeling in plants? *Plant science : an international journal of experimental plant biology,* 191-192**,** 53-70.

CORVAGLIA, A. R., FRANCOIS, P., HERNANDEZ, D., PERRON, K., LINDER, P. & SCHRENZEL, J. 2010. A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. *Proc Natl Acad Sci U S A,* 107**,** 11954-8.

COVERT, M. W., KNIGHT, E. M., REED, J. L., HERRGARD, M. J. & PALSSON, B. O. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature,* 429**,** 92-6.

CUI, L., MURAKAMI, H., KUWAHARA-ARAI, K., HANAKI, H. & HIRAMATSU, K. 2000. Contribution of a thickened cell wall and its glutamine nonamidated component to the vancomycin resistance expressed by Staphylococcus aureus Mu50. *Antimicrob Agents Chemother,* 44**,** 2276-85.

CVIJOVIC, M., BORDEL, S. & NIELSEN, J. 2011. Mathematical models of cell factories: moving towards the core of industrial biotechnology. *Microbial biotechnology,* 4**,** 572-84.

DAIRI, T. 2009. An alternative menaquinone biosynthetic pathway operating in microorganisms: an attractive target for drug discovery to pathogenic Helicobacter and Chlamydia strains. *J Antibiot (Tokyo),* 62**,** 347-52.

DAVIS, S. L., PERRI, M. B., DONABEDIAN, S. M., MANIERSKI, C., SINGH, A., VAGER, D., HAQUE, N. Z., SPEIRS, K., MUDER, R. R., ROBINSON-DUNN, B., HAYDEN, M. K. & ZERVOS, M. J. 2007. Epidemiology and outcomes of community-associated methicillin-resistant Staphylococcus aureus infection. *J Clin Microbiol,* 45**,** 1705-11.

DE CRECY-LAGARD, V., EL YACOUBI, B., DE LA GARZA, R. D., NOIRIEL, A. & HANSON, A. D. 2007. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics,* 8**,** 245.

DEBROY, C., ROBERTS, E. & FRATAMICO, P. M. 2011. Detection of O antigens in Escherichia coli. *Anim Health Res Rev,* 12**,** 169-85.

DELLOMONACO, C., CLOMBURG, J. M., MILLER, E. N. & GONZALEZ, R. 2011. Engineered reversal of the beta-oxidation cycle for the synthesis of fuels and chemicals. *Nature,* 476**,** 355-9.

DEMSAR, J., ZUPAN, B., LEBAN, G. & CURK, T. 2004. Orange: from experimental machine learning to interactive data mining. *Lecture notes in computer science*.

DENGER, K., WEISS, M., FELUX, A. K., SCHNEIDER, A., MAYER, C., SPITELLER, D., HUHN, T., COOK, A. M. & SCHLEHECK, D. 2014. Sulphoglycolysis in Escherichia coli K-12 closes a gap in the biogeochemical sulphur cycle. *Nature,* 507**,** 114-7.

DEVOID, S., OVERBEEK, R., DEJONGH, M., VONSTEIN, V., BEST, A. A. & HENRY, C. 2013. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods Mol Biol,* 985**,** 17-45.

DIAZ, E., FERRANDEZ, A., PRIETO, M. A. & GARCIA, J. L. 2001. Biodegradation of aromatic compounds by Escherichia coli. *Microbiol Mol Biol Rev,* 65**,** 523-69, table of contents.

DIEP, B. A., CHAMBERS, H. F., GRABER, C. J., SZUMOWSKI, J. D., MILLER, L. G., HAN, L. L., CHEN, J. H., LIN, F., LIN, J., PHAN, T. H., CARLETON, H. A., MCDOUGAL, L. K., TENOVER, F. C., COHEN, D. E., MAYER, K. H., SENSABAUGH, G. F. & PERDREAU-REMINGTON, F. 2008. Emergence of multidrug-resistant, community-associated, methicillin-resistant Staphylococcus aureus clone USA300 in men who have sex with men. *Ann Intern Med,* 148**,** 249-57.

DIEP, B. A., GILL, S. R., CHANG, R. F., PHAN, T. H., CHEN, J. H., DAVIDSON, M. G., LIN, F., LIN, J., CARLETON, H. A., MONGODIN, E. F., SENSABAUGH, G. F. & PERDREAU-REMINGTON, F. 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet,* 367**,** 731-9.

DONATI, C. & RAPPUOLI, R. 2013. Reverse vaccinology in the 21st century: improvements over the original design. *Ann N Y Acad Sci,* 1285**,** 115-32.

DUARTE, N. C., BECKER, S. A., JAMSHIDI, N., THIELE, I., MO, M. L., VO, T. D., SRIVAS, R. & PALSSON, B. O. 2007a. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 1777-82.

DUARTE, N. C., BECKER, S. A., JAMSHIDI, N., THIELE, I., MO, M. L., VO, T. D., SRIVAS, R. & PALSSON, B. O. 2007b. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A,* 104**,** 1777-82.

DURFEE, T., NELSON, R., BALDWIN, S., PLUNKETT, G., 3RD, BURLAND, V., MAU, B., PETROSINO, J. F., QIN, X., MUZNY, D. M., AYELE, M., GIBBS, R. A., CSORGO, B., POSFAI, G., WEINSTOCK, G. M. & BLATTNER, F. R. 2008. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *J Bacteriol,* 190**,** 2597-606.

DURINCK, S., MOREAU, Y., KASPRZYK, A., DAVIS, S., DE MOOR, B., BRAZMA, A. & HUBER, W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics,* 21**,** 3439-40.

EBRAHIM, A., ALMAAS, E., BAUER, E., BORDBAR, A., BURGARD, A. P., CHANG, R. L., DRAGER, A., FAMILI, I., FEIST, A. M., FLEMING, R. M., FONG, S. S., HATZIMANIKATIS, V., HERRGARD, M. J., HOLDER, A., HUCKA, M., HYDUKE, D., JAMSHIDI, N., LEE, S. Y., LE NOVERE, 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. 2015. Do genome-scale models need exact solvers or clearer standards? *Mol Syst Biol,* 11**,** 831.

EBRAHIM, A., LERMAN, J. A., PALSSON, B. O. & HYDUKE, D. R. 2013. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol,* 7**,** 74.

EDWARDS, J. S., IBARRA, R. U. & PALSSON, B. O. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology,* 19**,** 125-30.

EDWARDS, J. S. & PALSSON, B. O. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem,* 274**,** 17410-6.

EDWARDS, J. S. & PALSSON, B. O. 2000a. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A,* 97**,** 5528-33.

EDWARDS, J. S. & PALSSON, B. O. 2000b. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 5528-33.

ERBERSDOBLER, H. F. & FAIST, V. 2001. Metabolic transit of Amadori products. *Nahrung,* 45**,** 177-81.

FAINE, S. 1959. Iron as a growth requirement for pathogenic Leptospira. *J Gen Microbiol,* 20**,** 246-51.

FAINE, S. 1999. *Leptospira and Leptospirosis*, MediSci.

FAN, J., YE, J., KAMPHORST, J. J., SHLOMI, T., THOMPSON, C. B. & RABINOWITZ, J. D. 2014. Quantitative flux analysis reveals folate-dependent NADPH production. *Nature,* 510**,** 298-302.

FANG, G., ROCHA, E. & DANCHIN, A. 2005. How essential are nonessential genes? *Mol Biol Evol,* 22**,** 2147-56.

FEIL, E. J., COOPER, J. E., GRUNDMANN, H., ROBINSON, D. A., ENRIGHT, M. C., BERENDT, T., PEACOCK, S. J., SMITH, J. M., MURPHY, M., SPRATT, B. G., MOORE, C. E. & DAY, N. P. 2003. How clonal is Staphylococcus aureus? *J Bacteriol,* 185**,** 3307-16.

FEIST, A. M., HENRY, C. S., REED, J. L., KRUMMENACKER, M., JOYCE, A. R., KARP, P. D., BROADBELT, L. J., HATZIMANIKATIS, V. & PALSSON, B. O.

2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol,* 3**,** 121.

FEIST, A. M., HERRGARD, M. J., THIELE, I., REED, J. L. & PALSSON, B. O. 2009. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol,* 7**,** 129-43.

FEIST, A. M. & PALSSON, B. O. 2008a. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nature biotechnology,* 26**,** 659-67.

FEIST, A. M. & PALSSON, B. O. 2008b. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol,* 26**,** 659-67.

FEIST, A. M. & PALSSON, B. O. 2010a. The biomass objective function. *Current opinion in microbiology,* 13**,** 344-9.

FEIST, A. M. & PALSSON, B. O. 2010b. The biomass objective function. *Curr Opin Microbiol,* 13**,** 344-9.

FEIST, A. M. & PALSSON, B. O. 2010c. The biomass objective function. *Current Opinion in Microbiology,* 13**,** 344-349.

FITZGERALD, J. R., STURDEVANT, D. E., MACKIE, S. M., GILL, S. R. & MUSSER, J. M. 2001. Evolutionary genomics of Staphylococcus aureus: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci U S A,* 98**,** 8821-6.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., MERRICK, J. M. & ET AL. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269**,** 496-512.

FONG, N. L., LERMAN, J. A., LAM, I., PALSSON, B. O. & CHARUSANTI, P. 2013. Reconciling a Salmonella enterica metabolic model with experimental data confirms that overexpression of the glyoxylate shunt can rescue a lethal ppc deletion mutant. *FEMS Microbiol Lett,* 342**,** 62-9.

FONG, S. S., JOYCE, A. R. & PALSSON, B. O. 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome research,* 15**,** 1365-72.

FONG, S. S. & PALSSON, B. O. 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet,* 36**,** 1056-8.

FORSTER, J., FAMILI, I., FU, P., PALSSON, B. O. & NIELSEN, J. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res,* 13**,** 244-53.

FREDDOLINO, P. L., AMINI, S. & TAVAZOIE, S. 2012. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. *J Bacteriol,* 194**,** 303-6.

FRIDKIN, S. K., HAGEMAN, J., MCDOUGAL, L. K., MOHAMMED, J., JARVIS, W. R., PERL, T. M. & TENOVER, F. C. 2003. Epidemiological and microbiological characterization of infections caused by Staphylococcus aureus with reduced susceptibility to vancomycin, United States, 1997-2001. *Clin Infect Dis,* 36**,** 429-39.

FURUKAWA, H. & MIZUSHIMA, S. 1982. Roles of cell surface components of Escherichia coli K-12 in bacteriophage T4 infection: interaction of tail core with phospholipids. *J Bacteriol,* 150**,** 916-24.

GANTER, M., BERNARD, T., MORETTI, S., STELLING, J. & PAGNI, M. 2013. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics,* 29**,** 815-6.

GEER, L. Y., MARCHLER-BAUER, A., GEER, R. C., HAN, L., HE, J., HE, S., LIU, C., SHI, W. & BRYANT, S. H. 2010. The NCBI BioSystems database. *Nucleic Acids Res,* 38**,** D492-6.

GILIS, D., MASSAR, S., CERF, N. J. & ROOMAN, M. 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol,* 2**,** RESEARCH0049.

GIRONS, I. S., BOURHY, P., OTTONE, C., PICARDEAU, M., YELTON, D., HENDRIX, R. W., GLASER, P. & CHARON, N. 2000. The LE1 bacteriophage replicates as a plasmid within Leptospira biflexa: construction of an L. biflexa-Escherichia coli shuttle vector. *J Bacteriol,* 182**,** 5700-5.

GORDIENKO, E. N., KAZANOV, M. D. & GELFAND, M. S. 2013. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J Bacteriol,* 195**,** 2786-92.

GRAHAM, J. E. & WILKINSON, B. J. 1992. Staphylococcus aureus osmoregulation: roles for choline, glycine betaine, proline, and taurine. *J Bacteriol,* 174**,** 2711-6.

GROSJEAN, H., BRETON, M., SIRAND-PUGNET, P., TARDY, F., THIAUCOURT, F., CITTI, C., BARRE, A., YOSHIZAWA, S., FOURMY, D., DE CRECY-LAGARD, V. & BLANCHARD, A. 2014. Predicting the minimal translation apparatus: lessons from the reductive evolution of mollicutes. *PLoS Genet,* 10**,** e1004363.

GUEGAN, R., CAMADRO, J. M., SAINT GIRONS, I. & PICARDEAU, M. 2003. Leptospira spp. possess a complete haem biosynthetic pathway and are able to use exogenous haem sources. *Mol Microbiol,* 49**,** 745-54.

GULMEZIAN, M., HYMAN, K. R., MARBOIS, B. N., CLARKE, C. F. & JAVOR, G. T. 2007. The role of UbiX in Escherichia coli coenzyme Q biosynthesis. *Arch Biochem Biophys,* 467**,** 144-53.

GUROBI OPTIMIZATION, I. 2015. Gurobi Optimization, Inc.

GUTKNECHT, R., BEUTLER, R., GARCIA-ALLES, L. F., BAUMANN, U. & ERNI, B. 2001. The dihydroxyacetone kinase of Escherichia coli utilizes a phosphoprotein instead of ATP as phosphoryl donor. *EMBO J,* 20**,** 2480-6.

GUZMAN, G. I., UTRILLA, J., NURK, S., BRUNK, E., MONK, J. M., EBRAHIM, A., PALSSON, B. O. & FEIST, A. M. 2015. Model-driven discovery of underground metabolic functions in Escherichia coli. *Proc Natl Acad Sci U S A,* 112**,** 929-34.

HAASE, I., SARGE, S., ILLARIONOV, B., LAUDERT, D., HOHMANN, H. P., BACHER, A. & FISCHER, M. 2013. Enzymes from the haloacid dehalogenase (HAD) superfamily catalyse the elusive dephosphorylation step of riboflavin biosynthesis. *Chembiochem,* 14**,** 2272-5.

HABRYCH, M., RODRIGUEZ, S. & STEWART, J. D. 2002. Purification and identification of an Escherichia coli beta-keto ester reductase as 2,5-diketo-D-gluconate reductase YqhE. *Biotechnol Prog,* 18**,** 257-61.

HALL, B. G., EHRLICH, G. D. & HU, F. Z. 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology,* 156**,** 1060-8.

HANCOCK, R. 1960. The amino acid composition of the protein and cell wall of Staphylococcus aureus. *Biochim Biophys Acta,* 37**,** 42-6.

HANSEN, K. D., BRENNER, S. E. & DUDOIT, S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res,* 38**,** e131.

HARCOMBE, W. R., RIEHL, W. J., DUKOVSKI, I., GRANGER, B. R., BETTS, A., LANG, A. H., BONILLA, G., KAR, A., LEIBY, N., MEHTA, P., MARX, C. J. & SEGRE, D. 2014. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports,* 7**,** 1104-15.

HARWOOD, J. L. & NICHOLLS, R. G. 1979. The plant sulpholipid-- a major component of the sulphur cycle. *Biochem Soc Trans,* 7**,** 440-7.

HASSANINASAB, A., HASHIMOTO, Y., TOMITA-YOKOTANI, K. & KOBAYASHI, M. 2011. Discovery of the curcumin metabolic pathway involving a unique enzyme in an intestinal microorganism. *Proc Natl Acad Sci U S A,* 108**,** 6615-20.

HAYASHI, T., MAKINO, K., OHNISHI, M., KUROKAWA, K., ISHII, K., YOKOYAMA, K., HAN, C. G., OHTSUBO, E., NAKAYAMA, K., MURATA, T., TANAKA, M., TOBE, T., IIDA, T., TAKAMI, H., HONDA, T., SASAKAWA, C., OGASAWARA, N., YASUNAGA, T., KUHARA, S., SHIBA, T., HATTORI, M. & SHINAGAWA, H. 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res,* 8**,** 11-22.

HEAVNER, B. D. & PRICE, N. D. 2015. Transparency in metabolic network reconstruction enables scalable biological discovery. *Curr Opin Biotechnol,* 34**,** 105-9.

HEAVNER, B. D., SMALLBONE, K., BARKER, B., MENDES, P. & WALKER, L. P. 2012. Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. *BMC Syst Biol,* 6**,** 55.

HEINEMANN, M., KUMMEL, A., RUINATSCHA, R. & PANKE, S. 2005. In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. *Biotechnol Bioeng,* 92**,** 850-64.

HELDT, D., LAWRENCE, A. D., LINDENMEYER, M., DEERY, E., HEATHCOTE, P., RIGBY, S. E. & WARREN, M. J. 2005. Aerobic synthesis of vitamin B12: ring contraction and cobalt chelation. *Biochem Soc Trans,* 33**,** 815-9.

HENRY, C. S., BROADBELT, L. J. & HATZIMANIKATIS, V. 2007. Thermodynamics-based metabolic flux analysis. *Biophysical journal,* 92**,** 1792-805.

HENRY, C. S., DEJONGH, M., BEST, A. A., FRYBARGER, P. M., LINSAY, B. & STEVENS, R. L. 2010a. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology,* 28**,** 977-82.

HENRY, C. S., DEJONGH, M., BEST, A. A., FRYBARGER, P. M., LINSAY, B. & STEVENS, R. L. 2010b. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech,* 28**,** 977-982.

HENRY, C. S., DEJONGH, M., BEST, A. A., FRYBARGER, P. M., LINSAY, B. & STEVENS, R. L. 2010c. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol,* 28**,** 977-82.

HEROLD, B. C., IMMERGLUCK, L. C., MARANAN, M. C., LAUDERDALE, D. S., GASKIN, R. E., BOYLE-VAVRA, S., LEITCH, C. D. & DAUM, R. S. 1998. Community-acquired methicillin-resistant Staphylococcus aureus in children with no identified predisposing risk. *JAMA,* 279**,** 593-8.

HERRGARD, M. J., SWAINSTON, N., DOBSON, P., DUNN, W. B., ARGA, K. Y., ARVAS, M., BLUTHGEN, N., BORGER, S., COSTENOBLE, R., HEINEMANN, M., HUCKA, M., LE NOVERE, N., LI, P., LIEBERMEISTER, W., MO, M. L., OLIVEIRA, A. P., PETRANOVIC, D., PETTIFER, S., SIMEONIDIS, E., SMALLBONE, K., SPASIC, I., WEICHART, D., BRENT, R., BROOMHEAD, D.

S., WESTERHOFF, H. V., KIRDAR, B., PENTTILA, M., KLIPP, E., PALSSON, B. O., SAUER, U., OLIVER, S. G., MENDES, P., NIELSEN, J. & KELL, D. B. 2008. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol,* 26**,** 1155-60.

HIGHLANDER, S. K., HULTEN, K. G., QIN, X., JIANG, H., YERRAPRAGADA, S., MASON, E. O., JR., SHANG, Y., WILLIAMS, T. M., FORTUNOV, R. M., LIU, Y., IGBOELI, O., PETROSINO, J., TIRUMALAI, M., UZMAN, A., FOX, G. E., CARDENAS, A. M., MUZNY, D. M., HEMPHILL, L., DING, Y., DUGAN, S., BLYTH, P. R., BUHAY, C. J., DINH, H. H., HAWES, A. C., HOLDER, M., KOVAR, C. L., LEE, S. L., LIU, W., NAZARETH, L. V., WANG, Q., ZHOU, J., KAPLAN, S. L. & WEINSTOCK, G. M. 2007. Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus. *BMC Microbiol,* 7**,** 99.

HINDLE, E. 1925. Leptospira in London Waters. *Br Med J,* 2**,** 57-9.

HIRAMATSU, K., ARITAKA, N., HANAKI, H., KAWASAKI, S., HOSODA, Y., HORI, S., FUKUCHI, Y. & KOBAYASHI, I. 1997. Dissemination in Japanese hospitals of strains of Staphylococcus aureus heterogeneously resistant to vancomycin. *Lancet,* 350**,** 1670-3.

HIRATSUKA, T., FURIHATA, K., ISHIKAWA, J., YAMASHITA, H., ITOH, N., SETO, H. & DAIRI, T. 2008. An alternative menaquinone biosynthetic pathway operating in microorganisms. *Science,* 321**,** 1670-3.

HOLM, A. K., BLANK, L. M., OLDIGES, M., SCHMID, A., SOLEM, C., JENSEN, P. R. & VEMURI, G. N. 2010. Metabolic and transcriptional response to cofactor perturbations in Escherichia coli. *J Biol Chem,* 285**,** 17498-506.

HOLT, D. C., HOLDEN, M. T., TONG, S. Y., CASTILLO-RAMIREZ, S., CLARKE, L., QUAIL, M. A., CURRIE, B. J., PARKHILL, J., BENTLEY, S. D., FEIL, E. J. & GIFFARD, P. M. 2011. A very early-branching Staphylococcus aureus lineage lacking the carotenoid pigment staphyloxanthin. *Genome Biol Evol,* 3**,** 881-95.

HORMOZ, S. 2013. Amino acid composition of proteins reduces deleterious impact of mutations. *Sci Rep,* 3**,** 2919.

HTTP://SYSTEMSBIOLOGY.UCSD.EDU/INSILICOORGANISMS/OTHERORGANISMS.

HU, P., JANGA, S. C., BABU, M., DÍAZ-MEJÍA, J. J., BUTLAND, G., YANG, W., POGOUTSE, O., GUO, X., PHANSE, S., WONG, P., CHANDRAN, S., CHRISTOPOULOS, C., NAZARIANS-ARMAVIL, A., NASSERI, N. K., MUSSO, G., ALI, M., NAZEMOF, N., EROUKOVA, V., GOLSHANI, A., PACCANARO, A., GREENBLATT, J. F., MORENO-HAGELSIEB, G. & EMILI, A. 2009. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol,* 7**,** e1000096.

HUDSON, A. O., SINGH, B. K., LEUSTEK, T. & GILVARG, C. 2006. An LL-diaminopimelate aminotransferase defines a novel variant of the lysine biosynthesis pathway in plants. *Plant Physiol,* 140**,** 292-301.

HUERTA, A. M., SALGADO, H., THIEFFRY, D. & COLLADO-VIDES, J. 1998. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res,* 2**6,** 55-9.

HUGO, W. B. & DAVIDSON, J. R. 1973a. Effect of cell lipid depletion in Staphylococcus aureus upon its resistance to antimicrobial agents. 1. Lipid depletion induced by biotin deficiency. *Microbios,* 8**,** 43-51.

HUGO, W. B. & DAVIDSON, J. R. 1973b. Effect of cell lipid depletion in Staphylococcus aureus upon its resistance to antimicrobial agents. 2. A ccomparison of the response of normal and lipid depleted cells of S. aureus to antibacterial drugs. *Microbios,* 8**,** 63-72.

HUO, Y. X., CHO, K. M., RIVERA, J. G., MONTE, E., SHEN, C. R., YAN, Y. & LIAO, J. C. 2011. Conversion of proteins into biofuels by engineering nitrogen flux. *Nat Biotechnol,* 29**,** 346-51.

HUSSEIN, M. J., GREEN, J. M. & NICHOLS, B. P. 1998. Characterization of mutations that allow p-aminobenzoyl-glutamate utilization by Escherichia coli. *J Bacteriol,* 180**,** 6260-8.

HYDUKE, D. R., LEWIS, N. E. & PALSSON, B. O. 2013. Analysis of omics data with genome-scale models of metabolism. *Mol Biosyst,* 9**,** 167-74.

IBARRA, R. U., EDWARDS, J. S. & PALSSON, B. O. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature,* 420**,** 186-9.

ISHII, N., NAKAHIGASHI, K., BABA, T., ROBERT, M., SOGA, T., KANAI, A., HIRASAWA, T., NABA, M., HIRAI, K., HOQUE, A., HO, P. Y., KAKAZU, Y., SUGAWARA, K., IGARASHI, S., HARADA, S., MASUDA, T., SUGIYAMA, N., TOGASHI, T., HASEGAWA, M., TAKAI, Y., YUGI, K., ARAKAWA, K., IWATA, N., TOYA, Y., NAKAYAMA, Y., NISHIOKA, T., SHIMIZU, K., MORI, H. & TOMITA, M. 2007. Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science,* 316**,** 593-7.

JANG, Y. S., KIM, B., SHIN, J. H., CHOI, Y. J., CHOI, S., SONG, C. W., LEE, J., PARK, H. G. & LEE, S. Y. 2012. Bio-based production of C2-C6 platform chemicals. *Biotechnol Bioeng,* 109**,** 2437-59.

JANSEN, K. U., GIRGENTI, D. Q., SCULLY, I. L. & ANDERSON, A. S. 2013. Vaccine review: "Staphyloccocus aureus vaccines: problems and prospects". *Vaccine,* 31**,** 2723-30.

JANSSEN, P., GOLDOVSKY, L., KUNIN, V., DARZENTAS, N. & OUZOUNIS, C. A. 2005. Genome coverage, literally speaking. *EMBO Rep,* 6**,** 397-399.

JAUREGUY, F., LANDRAUD, L., PASSET, V., DIANCOURT, L., FRAPY, E., GUIGON, G., CARBONNELLE, E., LORTHOLARY, O., CLERMONT, O., DENAMUR, E., PICARD, B., NASSIF, X. & BRISSE, S. 2008. Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains. *BMC Genomics,* 9**,** 560.

JENKINS, L. S. & NUNN, W. D. 1987. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in Escherichia coli: the ato system. *J Bacteriol,* 169**,** 42-52.

JERBY, L., SHLOMI, T. & RUPPIN, E. 2010. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology,* 6**,** 401.

JNES E, OLIPHANT E & PETERSON P, E. A. 2001. *SciPy: Open Source Scientific Tools for Python* [Online]. Available: http://www.scipy.org/ [Accessed 2015-04-26].

JOHNSON, R. C. & GARY, N. D. 1962. Nutrition of Leptospira pomona. 1. A chemically defined substitute for rabbit serum ultrafiltrate. *J Bacteriol,* 83**,** 668-72.

JOHNSON, R. C., HARRIS, V. G. & WALBY, J. K. 1969. Characterization of leptospires according to fatty acid requirements. *J Gen Microbiol,* 55**,** 399-407.

JOHNSON, R. C., LIVERMORE, B. P., WALBY, J. K. & JENKIN, H. M. 1970. Lipids of parasitic and saprophytic leptospires. *Infect Immun,* 2**,** 286-91.

JOHNSON, R. C. & ROGERS, P. 1964. Differentiation of Pathogenic and Saprophytic Leptospires with 8-Azaguanine. *J Bacteriol,* 88**,** 1618-23.

KADIS, S. & PUGH, W. L. 1974. Urea utilization by Leptospira. *Infect Immun,* 10**,** 793-801.

KAMAT, S. S., WILLIAMS, H. J. & RAUSHEL, F. M. 2011. Intermediates in the transformation of phosphonates to phosphate by bacteria. *Nature,* 480**,** 570-3.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. & HIRAKAWA, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res,* 38**,** D355-60.

KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M. & TANABE, M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res,* 40**,** D109-14.

KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res,* 42**,** D199-205.

KARCH, H., TARR, P. I. & BIELASZEWSKA, M. 2005. Enterohaemorrhagic Escherichia coli in human medicine. *Int J Med Microbiol,* 295**,** 405-18.

KARR, J. R., SANGHVI, J. C., MACKLIN, D. N., GUTSCHOW, M. V., JACOBS, J. M., BOLIVAL, B., JR., ASSAD-GARCIA, N., GLASS, J. I. & COVERT, M. W. 2012. A whole-cell computational model predicts phenotype from genotype. *Cell,* 150**,** 389-401.

KASHKET, E. R. 1981. Proton motive force in growing Streptococcus lactis and Staphylococcus aureus cells under aerobic and anaerobic conditions. *J Bacteriol,* 146**,** 369-76.

KEFFORD, B. & MARSHALL, K. C. 1984. Adhesion of Leptospira at a solid-liquid interface: a model. *Arch Microbiol,* 138**,** 84-8.

KEHL-FIE, T. E. & SKAAR, E. P. 2010. Nutritional immunity beyond iron: a role for manganese and zinc. *Curr Opin Chem Biol,* 14**,** 218-24.

KESELER, I. M., BONAVIDES-MARTINEZ, C., COLLADO-VIDES, J., GAMA-CASTRO, S., GUNSALUS, R. P., JOHNSON, D. A., KRUMMENACKER, M., NOLAN, L. M., PALEY, S., PAULSEN, I. T., PERALTA-GIL, M., SANTOS-ZAVALETA, A., SHEARER, A. G. & KARP, P. D. 2009. EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res,* 37**,** D464-70.

KESELER, I. M., MACKIE, A., PERALTA-GIL, M., SANTOS-ZAVALETA, A., GAMA-CASTRO, S., BONAVIDES-MARTINEZ, C., FULCHER, C., HUERTA, A. M., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUNIZ-RASCADO, L., ONG, Q., PALEY, S., SCHRODER, I., SHEARER, A. G., SUBHRAVETI, P., TRAVERS, M., WEERASINGHE, D., WEISS, V., COLLADO-VIDES, J., GUNSALUS, R. P., PAULSEN, I. & KARP, P. D. 2013. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res,* 41**,** D605-12.

KHISAMOV, G. Z. & MOROZOVA, N. K. 1988. Fatty acids as resource of carbon for leptospirae. *J Hyg Epidemiol Microbiol Immunol,* 32**,** 87-93.

KIM, B., KIM, W. J., KIM, D. I. & LEE, S. Y. 2015. Applications of genome-scale metabolic network model in metabolic engineering. *J Ind Microbiol Biotechnol,* 42**,** 339-48.

KIM, B., PARK, H., NA, D. & LEE, S. Y. 2014. Metabolic engineering of Escherichia coli for the production of phenol from glucose. *Biotechnol J,* 9**,** 621-9.

KIM, H. U., KIM, S. Y., JEONG, H., KIM, T. Y., KIM, J. J., CHOY, H. E., YI, K. Y., RHEE, J. H. & LEE, S. Y. 2011. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Molecular systems biology,* 7**,** 460.

KIM, H. U., KIM, T. Y. & LEE, S. Y. 2010a. Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen Acinetobacter baumannii AYE. *Mol Biosyst,* 6**,** 339-48.

KIM, P. J., LEE, D. Y., KIM, T. Y., LEE, K. H., JEONG, H., LEE, S. Y. & PARK, S. 2007. Metabolite essentiality elucidates robustness of Escherichia coli metabolism. *Proc Natl Acad Sci U S A,* 104**,** 13638-42.

KIM, S. H. & LEE, P. C. 2012. Functional expression and extension of staphylococcal staphyloxanthin biosynthetic pathway in Escherichia coli. *J Biol Chem,* 287**,** 21575-83.

KIM, T. Y., KIM, H. U. & LEE, S. Y. 2010b. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab Eng,* 12**,** 105-11.

KIM, T. Y., SOHN, S. B., KIM, Y. B., KIM, W. J. & LEE, S. Y. 2012. Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr Opin Biotechnol,* 23**,** 617-623.

KING, Z. A., DRAGER, A., EBRAHIM, A., SONNENSCHEIN, N., LEWIS, N. E. & PALSSON, B. O. 2015a. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput Biol,* 11**,** e1004321.

KING, Z. A. & EBRAHIM, A. 2014. escher: Escher 1.0.0 Beta 3. *ZENODO*.

KING, Z. A., LLOYD, C. J., FEIST, A. M. & PALSSON, B. O. 2015b. Next-generation genome-scale models for metabolic engineering. *Current opinion in biotechnology,* 35C**,** 23-29.

KING, Z. A., LU, J., DRAGER, A., MILLER, P., FEDEROWICZ, S., LERMAN, J. A., EBRAHIM, A., PALSSON, B. O. & LEWIS, N. E. 2015c. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*.

KLEVENS, R. M., MORRISON, M. A., NADLE, J., PETIT, S., GERSHMAN, K., RAY, S., HARRISON, L. H., LYNFIELD, R., DUMYATI, G., TOWNES, J. M., CRAIG, A. S., ZELL, E. R., FOSHEIM, G. E., MCDOUGAL, L. K., CAREY, R. B. & FRIDKIN, S. K. 2007. Invasive methicillin-resistant Staphylococcus aureus infections in the United States. *JAMA,* 298**,** 1763-71.

KLITGORD, N. & SEGRE, D. 2010. Environments that induce synthetic microbial ecosystems. *PLoS computational biology,* 6**,** e1001002.

KLUYTMANS, J., VAN BELKUM, A. & VERBRUGH, H. 1997. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. *Clin Microbiol Rev,* 10**,** 505-20.

KOCH, H. U., HAAS, R. & FISCHER, W. 1984. The role of lipoteichoic acid biosynthesis in membrane lipid metabolism of growing Staphylococcus aureus. *Eur J Biochem,* 138**,** 357-63.

KOONIN, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol,* 1**,** 127-36.

KUMAR, V. S. & MARANAS, C. D. 2009. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol,* 5**,** e1000308.

KURODA, M., KURODA, H., OSHIMA, T., TAKEUCHI, F., MORI, H. & HIRAMATSU, K. 2003. Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in Staphylococcus aureus. *Mol Microbiol,* 49**,** 807-21.

KUZNETSOVA, E., PROUDFOOT, M., GONZALEZ, C. F., BROWN, G., OMELCHENKO, M. V., BOROZAN, I., CARMEL, L., WOLF, Y. I., MORI, H., SAVCHENKO, A. V., ARROWSMITH, C. H., KOONIN, E. V., EDWARDS, A. M. & YAKUNIN, A. F. 2006. Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. *J Biol Chem,* 281**,** 36149-61.

LANGILLE, M. G., HSIAO, W. W. & BRINKMAN, F. S. 2008. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics,* 9**,** 329.

LARSEN, A. R., BOCHER, S., STEGGER, M., GOERING, R., PALLESEN, L. V. & SKOV, R. 2008. Epidemiology of European community-associated methicillin-resistant Staphylococcus aureus clonal complex 80 type IV strains isolated in Denmark from 1993 to 2004. *J Clin Microbiol,* 46**,** 62-8.

LATIF, H., SZUBIN, R., TAN, J., BRUNK, E., LECHNER, A., ZENGLER, K. & PALSSON, B. O. 2015. A streamlined ribosome profiling protocol for the characterization of microorganisms. *Biotechniques,* 58**,** 329-32.

LAU, C. L., SMYTHE, L. D., CRAIG, S. B. & WEINSTEIN, P. 2010. Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? *Trans R Soc Trop Med Hyg,* 104**,** 631-8.

LAUPLAND, K. B., ROSS, T. & GREGSON, D. B. 2008. Staphylococcus aureus bloodstream infections: risk factors, outcomes, and the influence of methicillin resistance in Calgary, Canada, 2000-2006. *J Infect Dis,* 198**,** 336-43.

LAZZARONI, J. C., GERMON, P., RAY, M. C. & VIANNEY, A. 1999. The Tol proteins of Escherichia coli and their involvement in the uptake of biomolecules and outer membrane stability. *FEMS Microbiol Lett,* 177**,** 191-7.

LECOMPTE, O., RIPP, R., THIERRY, J. C., MORAS, D. & POCH, O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res,* 30**,** 5382-90.

LEE, D. H. & PALSSON, B. O. 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microbiol,* 76**,** 4158-68.

LEE, D. S., BURD, H., LIU, J., ALMAAS, E., WIEST, O., BARABASI, A. L., OLTVAI, Z. N. & KAPATRAL, V. 2009. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J Bacteriol,* 191**,** 4015-24.

LEE, J. W., NA, D., PARK, J. M., LEE, J., CHOI, S. & LEE, S. Y. 2012a. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat Chem Biol,* 8**,** 536-46.

LEE, K. H., PARK, J. H., KIM, T. Y., KIM, H. U. & LEE, S. Y. 2007. Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol,* 3**,** 149.

LEE, S. Y. 2009. *Systems Biology and Biotechnology of Escherichia Coli*, Springer.

LEE, S. Y., MATTANOVICH, D. & VILLAVERDE, A. 2012b. Systems metabolic engineering, industrial biotechnology and microbial cell factories. *Microb Cell Fact,* 11**,** 156.

LENNEN, R. M. & HERRGARD, M. J. 2014. Combinatorial strategies for improving multiple-stress resistance in industrially relevant Escherichia coli strains. *Appl Environ Microbiol,* 80**,** 6223-42.

LERMAN, J. A., HYDUKE, D. R., LATIF, H., PORTNOY, V. A., LEWIS, N. E., ORTH, J. D., SCHRIMPE-RUTLEDGE, A. C., SMITH, R. D., ADKINS, J. N., ZENGLER, K. & PALSSON, B. O. 2012a. In silico method for modelling metabolism and gene product expression at genome scale. *Nature communications,* 3**,** 929.

LERMAN, J. A., HYDUKE, D. R., LATIF, H., PORTNOY, V. A., LEWIS, N. E., ORTH, J. D., SCHRIMPE-RUTLEDGE, A. C., SMITH, R. D., ADKINS, J. N., ZENGLER, K. & PALSSON, B. O. 2012b. In silico method for modelling metabolism and gene product expression at genome scale. *Nat Commun,* 3**,** 929.

LESPINET, O. & LABEDAN, B. 2005. Orphan enzymes? *Science,* 307**,** 42.

LEVY, R. & BORENSTEIN, E. 2013. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 12804-9.

LEWIS, N. E., HIXSON, K. K., CONRAD, T. M., LERMAN, J. A., CHARUSANTI, P., POLPITIYA, A. D., ADKINS, J. N., SCHRAMM, G., PURVINE, S. O., LOPEZ-FERRER, D., WEITZ, K. K., EILS, R., KONIG, R., SMITH, R. D. & PALSSON, B. O. 2010a. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology,* 6**,** 390.

LEWIS, N. E., NAGARAJAN, H. & PALSSON, B. O. 2012a. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol,* 10**,** 291-305.

LEWIS, N. E., NAGARAJAN, H. & PALSSON, B. O. 2012b. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Micro,* 10**,** 291-305.

LEWIS, N. E., SCHRAMM, G., BORDBAR, A., SCHELLENBERGER, J., ANDERSEN, M. P., CHENG, J. K., PATEL, N., YEE, A., LEWIS, R. A., EILS, R., KONIG, R. & PALSSON, B. O. 2010b. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nature biotechnology,* 28**,** 1279-85.

LI, G. W., BURKHARDT, D., GROSS, C. & WEISSMAN, J. S. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell,* 157**,** 624-35.

LI, X., GIANOULIS, T. A., YIP, K. Y., GERSTEIN, M. & SNYDER, M. 2010. Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell,* 143**,** 639-50.

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. 2011a. An Experimentally Validated Genome-Scale Metabolic Reconstruction of *Klebsiella pneumoniae* MGH 78578, iYL1228. *J Bacteriol,* 193**,** 1710-1717.

LIAO, Y. C., HUANG, T. W., CHEN, F. C., CHARUSANTI, P., HONG, J. S., CHANG, H. Y., TSAI, S. F., PALSSON, B. O. & HSIUNG, C. A. 2011b. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J Bacteriol,* 193**,** 1710-7.

LINSTER, C. L., VAN SCHAFTINGEN, E. & HANSON, A. D. 2013. Metabolite damage and its repair or pre-emption. *Nat Chem Biol,* 9**,** 72-80.

LIU, G. Y., ESSEX, A., BUCHANAN, J. T., DATTA, V., HOFFMAN, H. M., BASTIAN, J. F., FIERER, J. & NIZET, V. 2005. Staphylococcus aureus golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. *J Exp Med,* 202**,** 209-15.

LIU, J. K., O'BRIEN, E. J., LERMAN, J. A., ZENGLER, K., PALSSON, B. O. & FEIST, A. M. 2014a. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Syst Biol,* 8**,** 110.

LIU, J. K., O'BRIEN, E. J., LERMAN, J. A., ZENGLER, K., PALSSON, B. O. & M., F. A. 2014b. Reconstruction and modeling protein translocation and

compartmentalization in Escherichia coli at the genome-scale. *BMC Systems Biology*.

LIU, X., LIU, L., WANG, Y., WANG, X., MA, Y. & LI, Y. 2014c. The Study on the factors affecting transformation efficiency of E. coli competent cells. *Pak J Pharm Sci,* 27**,** 679-84.

LOUVEL, H., BOMMEZZADRI, S., ZIDANE, N., BOURSAUX-EUDE, C., CRENO, S., MAGNIER, A., ROUY, Z., MEDIGUE, C., SAINT GIRONS, I., BOUCHIER, C. & PICARDEAU, M. 2006. Comparative and functional genomic analyses of iron transport and regulation in Leptospira spp. *J Bacteriol,* 188**,** 7893-904.

LOVELL, R. & HARVEY, D. G. 1950. A preliminary study of ammonia production by Corynebacterium renale and some other pathogenic bacteria. *J Gen Microbiol,* 4**,** 493-500.

LUKJANCENKO, O., WASSENAAR, T. M. & USSERY, D. W. 2010. Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol,* 60**,** 708-20.

MACHADO, D. & HERRGARD, M. 2014. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol,* 10**,** e1003580.

MAHADEVAN, R. & SCHILLING, C. H. 2003a. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng,* 5**,** 264-76.

MAHADEVAN, R. & SCHILLING, C. H. 2003b. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering,* 5**,** 264-76.

MARBAIX, A. Y., NOEL, G., DETROUX, A. M., VERTOMMEN, D., VAN SCHAFTINGEN, E. & LINSTER, C. L. 2011. Extremely conserved ATP- or ADP-dependent enzymatic system for nicotinamide nucleotide repair. *J Biol Chem,* 286**,** 41246-52.

MAREE, C. L., DAUM, R. S., BOYLE-VAVRA, S., MATAYOSHI, K. & MILLER, L. G. 2007. Community-associated methicillin-resistant Staphylococcus aureus isolates causing healthcare-associated infections. *Emerg Infect Dis,* 13**,** 236-42.

MARISCH, K., BAYER, K., SCHARL, T., MAIRHOFER, J., KREMPL, P. M., HUMMEL, K., RAZZAZI-FAZELI, E. & STRIEDNER, G. 2013. A comparative analysis of industrial Escherichia coli K-12 and B strains in high-glucose batch cultivations on process-, transcriptome- and proteome level. *PLoS One,* 8**,** e70516.

MARQUES, J. C., LAMOSA, P., RUSSELL, C., VENTURA, R., MAYCOCK, C., SEMMELHACK, M. F., MILLER, S. T. & XAVIER, K. B. 2011. Processing the interspecies quorum-sensing signal autoinducer-2 (AI-2): characterization of phospho-(S)-4,5-dihydroxy-2,3-pentanedione isomerization by LsrG protein. *J Biol Chem,* 286**,** 18331-43.

MASSEY, V. 1994. Activation of molecular oxygen by flavins and flavoproteins. *J Biol Chem,* 269**,** 22459-62.

MATTHIAS, M. A., RICALDI, J. N., CESPEDES, M., DIAZ, M. M., GALLOWAY, R. L., SAITO, M., STEIGERWALT, A. G., PATRA, K. P., ORE, C. V., GOTUZZO, E., GILMAN, R. H., LEVETT, P. N. & VINETZ, J. M. 2008. Human leptospirosis caused by a new, antigenically unique Leptospira associated with a Rattus species reservoir in the Peruvian Amazon. *PLoS Negl Trop Dis,* 2**,** e213.

MCCARTHY, A. J. & LINDSAY, J. A. 2012. The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC Microbiol,* 12**,** 104.

MCCLOSKEY, D., PALSSON, B. O. & FEIST, A. M. 2013a. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Molecular systems biology,* 9**,** 661.

MCCLOSKEY, D., PALSSON, B. O. & FEIST, A. M. 2013b. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol,* 9**,** 661.

MCCLOSKEY, D., PALSSON, B. O. & FEIST, A. M. 2013c. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Molecular Systems Biology*.

MCCLOSKEY, D., PALSSON, B. O. & FEIST, A. M. 2013d. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol Syst Biol,* 9.

MCKINNEY, W. Year. pandas: a Foundational Python Library for Data Analysis and Statistics. *In:* PyHPC2011, 2011.

MCNAMARA, P. J. & PROCTOR, R. A. 2000. Staphylococcus aureus small colony variants, electron transport and persistent infections. *Int J Antimicrob Agents,* 14**,** 117-22.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* 15**,** 589-94.

MEGCHELENBRINK, W., HUYNEN, M. & MARCHIORI, E. 2014. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS One,* 9**,** e86587.

MEIER, S., JENSEN, P. R. & DUUS, J. O. 2012. Direct observation of metabolic differences in living Escherichia coli strains K-12 and BL21. *Chembiochem,* 13**,** 308-10.

MESSNER, K. R. & IMLAY, J. A. 1999. The identification of primary sites of superoxide and hydrogen peroxide formation in the aerobic respiratory chain and sulfite reductase complex of Escherichia coli. *J Biol Chem,* 274**,** 10119-28.

MILLER, J. 1972. *Experiments in Molecular Genetics,* Cold Spring Harbor, NY, Cold Spring Harbor Laboratory.

MO, M. L., PALSSON, B. O. & HERRGARD, M. J. 2009. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology,* 3**,** 37.

MOLINA-HENARES, M. A., DE LA TORRE, J., GARCIA-SALAMANCA, A., MOLINA-HENARES, A. J., HERRERA, M. C., RAMOS, J. L. & DUQUE, E. 2010. Identification of conditionally essential genes for growth of Pseudomonas putida KT2440 on minimal medium through the screening of a genome-wide mutant library. *Environ Microbiol,* 12**,** 1468-85.

MONK, J., NOGALES, J. & PALSSON, B. O. 2014a. Optimizing genome-scale network reconstructions. *Nature biotechnology,* 32**,** 447-52.

MONK, J., NOGALES, J. & PALSSON, B. O. 2014b. Optimizing genome-scale network reconstructions. *Nat Biotechnol,* 32**,** 447-52.

MONK, J. & PALSSON, B. O. 2014. Genetics. Predicting microbial growth. *Science,* 344**,** 1448-9.

MONK, J. M., CHARUSANTI, P., AZIZ, R. K., LERMAN, J. A., PREMYODHIN, N., ORTH, J. D., FEIST, A. M. & PALSSON, B. O. 2013. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A,* 110**,** 20338-43.

MOORE, S. J. & WARREN, M. J. 2012. The anaerobic biosynthesis of vitamin B12. *Biochem Soc Trans,* 40**,** 581-6.

MORAN, G. J., KRISHNADASAN, A., GORWITZ, R. J., FOSHEIM, G. E., MCDOUGAL, L. K., CAREY, R. B. & TALAN, D. A. 2006. Methicillin-resistant S. aureus infections among patients in the emergency department. *N Engl J Med,* 355**,** 666-74.

MORRISON, C. 2015. Antibacterial antibodies gain traction. *Nat Rev Drug Discov,* 14**,** 737-8.

MURACHI, T. & TABATA, M. 1987. Use of a bioreactor consisting of sequentially aligned L-glutamate dehydrogenase and L-glutamate oxidase for the determination of ammonia by chemiluminescence. *Biotechnol Appl Biochem,* 9**,** 303-9.

MURRAY, G. L., SRIKRAM, A., HENRY, R., PUAPAIROJ, A., SERMSWAN, R. W. & ADLER, B. 2009. Leptospira interrogans requires heme oxygenase for disease pathogenesis. *Microbes Infect,* 11**,** 311-4.

MUSHEGIAN, A. 2008. Gene content of LUCA, the last universal common ancestor. *Front Biosci,* 13**,** 4657-66.

NA, D., YOO, S. M., CHUNG, H., PARK, H., PARK, J. H. & LEE, S. Y. 2013. Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. *Nat Biotechnol,* 31**,** 170-4.

NAGARAJAN, H., EMBREE, M., ROTARU, A. E., SHRESTHA, P. M., FEIST, A. M., PALSSON, B. O., LOVLEY, D. R. & ZENGLER, K. 2013. Characterization and modelling of interspecies electron transfer mechanisms and microbial community dynamics of a syntrophic association. *Nature communications,* 4**,** 2809.

NAHKU, R., PEEBO, K., VALGEPEA, K., BARRICK, J. E., ADAMBERG, K. & VILU, R. 2011. Stock culture heterogeneity rather than new mutational variation complicates short-term cell physiology studies of Escherichia coli K-12 MG1655 in continuous culture. *Microbiology,* 157**,** 2604-10.

NAKAHIGASHI, K., TOYA, Y., ISHII, N., SOGA, T., HASEGAWA, M., WATANABE, H., TAKAI, Y., HONMA, M., MORI, H. & TOMITA, M. 2009a. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol,* 5**,** 306.

NAKAHIGASHI, K., TOYA, Y., ISHII, N., SOGA, T., HASEGAWA, M., WATANABE, H., TAKAI, Y., HONMA, M., MORI, H. & TOMITA, M. 2009b. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Molecular systems biology,* 5**,** 306.

NAKANO, M., IIDA, T., OHNISHI, M., KUROKAWA, K., TAKAHASHI, A., TSUKAMOTO, T., YASUNAGA, T., HAYASHI, T. & HONDA, T. 2001. Association of the urease gene with enterohemorrhagic Escherichia coli strains irrespective of their serogroups. *J Clin Microbiol,* 39**,** 4541-3.

NAM, H., LEWIS, N. E., LERMAN, J. A., LEE, D. H., CHANG, R. L., KIM, D. & PALSSON, B. O. 2012. Network context and selection in the evolution to enzyme specificity. *Science,* 337**,** 1101-4.

NANRA, J. S., BUITRAGO, S. M., CRAWFORD, S., NG, J., FINK, P. S., HAWKINS, J., SCULLY, I. L., MCNEIL, L. K., ASTE-AMEZAGA, J. M., COOPER, D., JANSEN, K. U. & ANDERSON, A. S. 2013. Capsular polysaccharides are an important immune evasion mechanism for Staphylococcus aureus. *Hum Vaccin Immunother,* 9**,** 480-7.

NASCIMENTO, A. L., KO, A. I., MARTINS, E. A., MONTEIRO-VITORELLO, C. B., HO, P. L., HAAKE, D. A., VERJOVSKI-ALMEIDA, S., HARTSKEERL, R. A., MARQUES, M. V., OLIVEIRA, M. C., MENCK, C. F., LEITE, L. C., CARRER, H., COUTINHO, L. L., DEGRAVE, W. M., DELLAGOSTIN, O. A., EL-DORRY, H., FERRO, E. S., FERRO, M. I., FURLAN, L. R., GAMBERINI, M., GIGLIOTI, E. A., GOES-NETO, A., GOLDMAN, G. H., GOLDMAN, M. H., HARAKAVA, R., JERONIMO, S. M., JUNQUEIRA-DE-AZEVEDO, I. L., KIMURA, E. T.,

KURAMAE, E. E., LEMOS, E. G., LEMOS, M. V., MARINO, C. L., NUNES, L. R., DE OLIVEIRA, R. C., PEREIRA, G. G., REIS, M. S., SCHRIEFER, A., SIQUEIRA, W. J., SOMMER, P., TSAI, S. M., SIMPSON, A. J., FERRO, J. A., CAMARGO, L. E., KITAJIMA, J. P., SETUBAL, J. C. & VAN SLUYS, M. A. 2004a. Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol,* 186**,** 2164-72.

NASCIMENTO, A. L., VERJOVSKI-ALMEIDA, S., VAN SLUYS, M. A., MONTEIRO-VITORELLO, C. B., CAMARGO, L. E., DIGIAMPIETRI, L. A., HARSTKEERL, R. A., HO, P. L., MARQUES, M. V., OLIVEIRA, M. C., SETUBAL, J. C., HAAKE, D. A. & MARTINS, E. A. 2004b. Genome features of Leptospira interrogans serovar Copenhageni. *Braz J Med Biol Res,* 37**,** 459-77.

NEUHAUS, F. C. & BADDILEY, J. 2003. A continuum of anionic charge: structures and functions of D-alanyl-teichoic acids in gram-positive bacteria. *Microbiol Mol Biol Rev,* 67**,** 686-723.

NICHOLS, B. P. & GREEN, J. M. 1992. Cloning and sequencing of Escherichia coli ubiC and purification of chorismate lyase. *J Bacteriol,* 174**,** 5309-16.

NIZET, V. 2007. Understanding how leading bacterial pathogens subvert innate immunity to reveal novel therapeutic targets. *J Allergy Clin Immunol,* 120**,** 13-22.

NOBELMANN, B. & LENGELER, J. W. 1995. Sequence of the gat operon for galactitol utilization from a wild-type strain EC3132 of Escherichia coli. *Biochim Biophys Acta,* 1262**,** 69-72.

NOBELMANN, B. & LENGELER, J. W. 1996. Molecular analysis of the gat genes from Escherichia coli and of their roles in galactitol transport and metabolism. *J Bacteriol,* 178**,** 6790-5.

NOBLE, W. C., VIRANI, Z. & CREE, R. G. 1992. Co-transfer of vancomycin and other resistance genes from Enterococcus faecalis NCTC 12201 to Staphylococcus aureus. *FEMS Microbiol Lett,* 72**,** 195-8.

NOGALES, J., GUDMUNDSSON, S., KNIGHT, E. M., PALSSON, B. O. & THIELE, I. 2012. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc Natl Acad Sci U S A,* 109**,** 2678-2683.

NOINAJ, N., GUILLIER, M., BARNARD, T. J. & BUCHANAN, S. K. 2010. TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol,* 64**,** 43-60.

NONET, M. L., MARVEL, C. C. & TOLAN, D. R. 1987. The hisT-purF region of the Escherichia coli K-12 chromosome. Identification of additional genes of the hisT and purF operons. *J Biol Chem,* 262**,** 12209-17.

NOOR, E., LEWIS, N. E. & MILO, R. 2012. A proof for loop-law constraints in stoichiometric metabolic networks. *BMC systems biology,* 6**,** 140.

NOTEBAART, R. A., SZAPPANOS, B., KINTSES, B., PAL, F., GYORKEI, A., BOGOS, B., LAZAR, V., SPOHN, R., CSORGO, B., WAGNER, A., RUPPIN, E., PAL, C. & PAPP, B. 2014. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 11762-7.

O'BRIEN, E. J., LERMAN, J. A., CHANG, R. L., HYDUKE, D. R. & PALSSON, B. O. 2013a. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology,* 9**,** 693.

O'BRIEN, E. J., LERMAN, J. A., CHANG, R. L., HYDUKE, D. R. & PALSSON, B. O. 2013b. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol,* 9**,** 693.

O'BRIEN, E. J., MONK, J. M. & PALSSON, B. O. 2015. Using Genome-scale Models to Predict Biological Capabilities. *Cell,* 161**,** 971-87.

O'BRIEN, E. J. & PALSSON, B. O. 2015. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Current opinion in biotechnology,* 34C**,** 125-134.

OBERHARDT, M. A., PALSSON, B. O. & PAPIN, J. A. 2009. Applications of genome-scale metabolic reconstructions. *Molecular systems biology,* 5**,** 320.

OBERHARDT, M. A., PUCHALKA, J., MARTINS DOS SANTOS, V. A. & PAPIN, J. A. 2011. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS computational biology,* 7**,** e1001116.

ORTH, J., FLEMING, R. & PALSSON, B. O. 2010a. Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. *EcoSal -- Escherichia coli and Salmonella Cellular and Molecular Biology, Karp PD (ed).* 10.2.1 ed. Washington DC: ASM Press.

ORTH, J. D., CONRAD, T. M., NA, J., LERMAN, J. A., NAM, H., FEIST, A. M. & PALSSON, B. O. 2011. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol,* 7**,** 535.

ORTH, J. D. & PALSSON, B. 2012. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC Syst Biol,* 6**,** 30.

ORTH, J. D. & PALSSON, B. O. 2010a. Systematizing the generation of missing metabolic knowledge. *Biotechnol Bioeng,* 107**,** 403-12.

ORTH, J. D. & PALSSON, B. O. 2010b. Systematizing the generation of missing metabolic knowledge. *Biotechnology and bioengineering,* 107**,** 403-12.

ORTH, J. D., THIELE, I. & PALSSON, B. O. 2010b. What is flux balance analysis? *Nat Biotechnol,* 28**,** 245-8.

ORTH, J. D., THIELE, I. & PALSSON, B. O. 2010c. What is flux balance analysis? *Nature biotechnology,* 28**,** 245-8.

ÖSTERLUND, T., NOOKAEW, I. & NIELSEN, J. 2012. Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. *Biotechnol Adv* 30**,** 979-988.

PALSSON, B. 2000. The challenges of in silico biology. *Nature biotechnology,* 18**,** 1147-50.

PARK, B., NIZET, V. & LIU, G. Y. 2008. Role of Staphylococcus aureus catalase in niche competition against Streptococcus pneumoniae. *J Bacteriol,* 190**,** 2275-8.

PECHOUS, R., LEDALA, N., WILKINSON, B. J. & JAYASWAL, R. K. 2004. Regulation of the expression of cell wall stress stimulon member gene msrA1 in methicillin-susceptible or -resistant Staphylococcus aureus. *Antimicrob Agents Chemother,* 48**,** 3057-63.

PERNA, N. T., PLUNKETT, G., 3RD, BURLAND, V., MAU, B., GLASNER, J. D., ROSE, D. J., MAYHEW, G. F., EVANS, P. S., GREGOR, J., KIRKPATRICK, H. A., POSFAI, G., HACKETT, J., KLINK, S., BOUTIN, A., SHAO, Y., MILLER, L., GROTBECK, E. J., DAVIS, N. W., LIM, A., DIMALANTA, E. T., POTAMOUSIS, K. D., APODACA, J., ANANTHARAMAN, T. S., LIN, J., YEN, G., SCHWARTZ, D. C., WELCH, R. A. & BLATTNER, F. R. 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature,* 409**,** 529-33.

PICARDEAU, M., BRENOT, A. & SAINT GIRONS, I. 2001. First evidence for gene replacement in Leptospira spp. Inactivation of L. biflexa flaB results in non-motile mutants deficient in endoflagella. *Mol Microbiol,* 40**,** 189-99.

PICARDEAU, M., BULACH, D. M., BOUCHIER, C., ZUERNER, R. L., ZIDANE, N., WILSON, P. J., CRENO, S., KUCZEK, E. S., BOMMEZZADRI, S., DAVIS, J. C., MCGRATH, A., JOHNSON, M. J., BOURSAUX-EUDE, C., SEEMANN, T., ROUY, Z., COPPEL, R. L., ROOD, J. I., LAJUS, A., DAVIES, J. K., MEDIGUE, C. & ADLER, B. 2008. Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of Leptospira and the pathogenesis of leptospirosis. *PLoS One,* 3**,** e1607.

PINSKE, C., BONN, M., KRUGER, S., LINDENSTRAUSS, U. & SAWERS, R. G. 2011. Metabolic deficiences revealed in the biotechnologically important model bacterium Escherichia coli BL21(DE3). *PLoS One,* 6**,** e22830.

PLATA, G., HENRY, C. S. & VITKUP, D. 2015. Long-term phenotypic evolution of bacteria. *Nature,* 517**,** 369-72.

POULIOT, Y. & KARP, P. 2007. A survey of orphan enzyme activities. *BMC Bioinformatics,* 8**,** 244.

PRIBAT, A., JEANGUENIN, L., LARA-NUNEZ, A., ZIEMAK, M. J., HYDE, J. E., DE CRECY-LAGARD, V. & HANSON, A. D. 2009. 6-pyruvoyltetrahydropterin synthase paralogs replace the folate synthesis enzyme dihydroneopterin aldolase in diverse bacteria. *J Bacteriol,* 191**,** 4158-65.

PRICE, N. D., REED, J. L. & PALSSON, B. O. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol,* 2**,** 886-97.

PRIETO, M. A., DIAZ, E. & GARCIA, J. L. 1996. Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of Escherichia coli W: engineering a mobile aromatic degradative cluster. *J Bacteriol,* 178**,** 111-20.

PRUNIER, A. L., SCHUCH, R., FERNANDEZ, R. E. & MAURELLI, A. T. 2007a. Genetic structure of the nadA and nadB antivirulence loci in Shigella spp. *J Bacteriol,* 189**,** 6482-6.

PRUNIER, A. L., SCHUCH, R., FERNANDEZ, R. E., MUMY, K. L., KOHLER, H., MCCORMICK, B. A. & MAURELLI, A. T. 2007b. nadA and nadB of Shigella flexneri 5a are antivirulence loci responsible for the synthesis of quinolinate, a small molecule inhibitor of Shigella pathogenicity. *Microbiology,* 153**,** 2363-72.

PUPO, G. M., LAN, R. & REEVES, P. R. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A,* 97**,** 10567-72.

QU, J. 2007. *Sequencing and comparative genomics of Leptospira interrogans serovar pomona and Leptospira kirschneri serovar grippotyphosa.* Biomedical Engineering, University of Oklahoma.

RABINOWITZ, J. D. & VASTAG, L. 2012. Teaching the design principles of metabolism. *Nat Chem Biol,* 8**,** 497-501.

RAFTER, G. W., CHAYKIN, S. & KREBS, E. G. 1954. The action of glyceraldehyde-3-phosphate dehydrogenase on reduced diphosphopyridine nucleotide. *J Biol Chem,* 208**,** 799-811.

RANGANATHAN, S., SUTHERS, P. F. & MARANAS, C. D. 2010. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS computational biology,* 6**,** e1000744.

RATELADE, J., MIOT, M. C., JOHNSON, E., BETTON, J. M., MAZODIER, P. & BENAROUDJ, N. 2009. Production of recombinant proteins in the lon-deficient BL21(DE3) strain of Escherichia coli in the absence of the DnaK chaperone. *Appl Environ Microbiol,* 75**,** 3803-7.

RATET, G., VEYRIER, F. J., FANTON D'ANDON, M., KAMMERSCHEIT, X., NICOLA, M. A., PICARDEAU, M., BONECA, I. G. & WERTS, C. 2014. Live imaging of bioluminescent leptospira interrogans in mice reveals renal colonization as a

stealth escape from the blood defenses and antibiotics. *PLoS Negl Trop Dis,* 8**,** e3359.

RAUX, E., SCHUBERT, H. L. & WARREN, M. J. 2000. Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. *Cell Mol Life Sci,* 57**,** 1880-93.

RAVIKRISHNAN, A. & RAMAN, K. 2015. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Brief Bioinform,* 16**,** 1057-68.

REDDY, V. S. & SAIER, M. H., JR. 2012. BioV Suite--a collection of programs for the study of transport protein evolution. *FEBS J,* 279**,** 2036-46.

REED, J. L. 2012. Shrinking the metabolic solution space using experimental datasets. *PLoS computational biology,* 8**,** e1002662.

REED, J. L. & PALSSON, B. O. 2003a. Thirteen years of building constraint-based in silico models of Escherichia coli. *Journal of bacteriology,* 185**,** 2692-9.

REED, J. L. & PALSSON, B. Ø. 2003b. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol,* 185**,** 2692-2699.

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. 2006a. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A,* 103**,** 17480-4.

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. 2006b. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 17480-4.

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. Ø. 2006c. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A,* 103**,** 17480-4.

REED, J. L., VO, T. D., SCHILLING, C. H. & PALSSON, B. O. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol,* 4**,** R54.

REN, S. X., FU, G., JIANG, X. G., ZENG, R., MIAO, Y. G., XU, H., ZHANG, Y. X., XIONG, H., LU, G., LU, L. F., JIANG, H. Q., JIA, J., TU, Y. F., JIANG, J. X., GU, W. Y., ZHANG, Y. Q., CAI, Z., SHENG, H. H., YIN, H. F., ZHANG, Y., ZHU, G. F., WAN, M., HUANG, H. L., QIAN, Z., WANG, S. Y., MA, W., YAO, Z. J., SHEN, Y., QIANG, B. Q., XIA, Q. C., GUO, X. K., DANCHIN, A., SAINT GIRONS, I., SOMERVILLE, R. L., WEN, Y. M., SHI, M. H., CHEN, Z., XU, J. G. & ZHAO, G. P. 2003. Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature,* 422**,** 888-93.

REUSCH, V. M., JR. 1984. Lipopolymers, isoprenoids, and the assembly of the gram-positive cell wall. *Crit Rev Microbiol,* 11**,** 129-55.

RICALDI, J. N., FOUTS, D. E., SELENGUT, J. D., HARKINS, D. M., PATRA, K. P., MORENO, A., LEHMANN, J. S., PURUSHE, J., SANKA, R., TORRES, M., WEBSTER, N. J., VINETZ, J. M. & MATTHIAS, M. A. 2012. Whole genome analysis of Leptospira licerasiae provides insight into leptospiral evolution and pathogenicity. *PLoS Negl Trop Dis,* 6**,** e1853.

RILEY, M., ABE, T., ARNAUD, M. B., BERLYN, M. K. B., BLATTNER, F. R., CHAUDHURI, R. R., GLASNER, J. D., HORIUCHI, T., KESELER, I. M., KOSUGE, T., MORI, H., PERNA, N. T., PLUNKETT, G., RUDD, K. E., SERRES, M. H., THOMAS, G. H., THOMSON, N. R., WISHART, D. & WANNER, B. L. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Research,* 34**,** 1-9.

ROBICHON, C., LUO, J., CAUSEY, T. B., BENNER, J. S. & SAMUELSON, J. C. 2011. Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli protein contamination after purification by immobilized metal affinity chromatography. *Appl Environ Microbiol,* 77**,** 4634-46.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

RODIONOV, D. A., VITRESCHAK, A. G., MIRONOV, A. A. & GELFAND, M. S. 2003. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem,* 278**,** 41148-59.

ROLFSSON, O., PALSSON, B. O. & THIELE, I. 2011a. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC systems biology,* 5**,** 155.

ROLFSSON, O., PALSSON, B. O. & THIELE, I. 2011b. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst Biol,* 5**,** 155.

ROY, A., KUCUKURAL, A. & ZHANG, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc,* 5**,** 725-38.

ROY, A. B., HEWLINS, M. J., ELLIS, A. J., HARWOOD, J. L. & WHITE, G. F. 2003. Glycolytic breakdown of sulfoquinovose in bacteria: a missing link in the sulfur cycle. *Appl Environ Microbiol,* 69**,** 6434-41.

SAIER, M. H., JR., REDDY, V. S., TAMANG, D. G. & VASTERMARK, A. 2014. The transporter classification database. *Nucleic Acids Res,* 42**,** D251-8.

SAITO, M., MIYAHARA, S., VILLANUEVA, S. Y., ARAMAKI, N., IKEJIRI, M., KOBAYASHI, Y., GUEVARRA, J. P., MASUZAWA, T., GLORIANI, N. G.,

YANAGIHARA, Y. & YOSHIDA, S. 2014. PCR and culture identification of pathogenic Leptospira spp. from coastal soil in Leyte, Philippines, after a storm surge during Super Typhoon Haiyan (Yolanda). *Appl Environ Microbiol,* 80**,** 6926-32.

SATISH KUMAR, V., DASIKA, M. S. & MARANAS, C. D. 2007. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics,* 8**,** 212.

SCHEIBE, T. D., MAHADEVAN, R., FANG, Y., GARG, S., LONG, P. E. & LOVLEY, D. R. 2009. Coupling a genome-scale metabolic model with a reactive transport model to describe in situ uranium bioremediation. *Microbial biotechnology,* 2**,** 274-86.

SCHELLENBERGER, J., LEWIS, N. E. & PALSSON, B. O. 2011a. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical journal,* 100**,** 544-53.

SCHELLENBERGER, J. & PALSSON, B. O. 2009a. Use of randomized sampling for analysis of metabolic networks. *The Journal of biological chemistry,* 284**,** 5457-61.

SCHELLENBERGER, J. & PALSSON, B. O. 2009b. Use of randomized sampling for analysis of metabolic networks. *J Biol Chem,* 284**,** 5457-61.

SCHELLENBERGER, J., PARK, J. O., CONRAD, T. M. & PALSSON, B. O. 2010. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics,* 11**,** 213.

SCHELLENBERGER, J., QUE, R., FLEMING, R. M., THIELE, I., ORTH, J. D., FEIST, A. M., ZIELINSKI, D. C., BORDBAR, A., LEWIS, N. E., RAHMANIAN, S., KANG, J., HYDUKE, D. R. & PALSSON, B. O. 2011b. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols,* 6**,** 1290-307.

SCHELLENBERGER, J., QUE, R., FLEMING, R. M., THIELE, I., ORTH, J. D., FEIST, A. M., ZIELINSKI, D. C., BORDBAR, A., LEWIS, N. E., RAHMANIAN, S., KANG, J., HYDUKE, D. R. & PALSSON, B. O. 2011c. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc,* 6**,** 1290-307.

SCHLEIFER, K. H. & KANDLER, O. 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriol Rev,* 36**,** 407-77.

SCHMIDT, B. J., EBRAHIM, A., METZ, T. O., ADKINS, J. N., PALSSON, B. O. & HYDUKE, D. R. 2013a. GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics,* 29**,** 2900-8.

SCHMIDT, B. J., PAPIN, J. A. & MUSANTE, C. J. 2013b. Mechanistic systems modeling to guide drug discovery and development. *Drug Discov Today,* 18**,** 116-27.

SCHNOES, A. M., BROWN, S. D., DODEVSKI, I. & BABBITT, P. C. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol,* 5**,** e1000605.

SCHOMBURG, I., CHANG, A., PLACZEK, S., SOHNGEN, C., ROTHER, M., LANG, M., MUNARETTO, C., ULAS, S., STELZER, M., GROTE, A., SCHEER, M. & SCHOMBURG, D. 2013. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res,* 41**,** D764-72.

SCHUETZ, R., KUEPFER, L. & SAUER, U. 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular systems biology,* 3**,** 119.

SCHUETZ, R., ZAMBONI, N., ZAMPIERI, M., HEINEMANN, M. & SAUER, U. 2012. Multidimensional optimality of microbial metabolism. *Science,* 336**,** 601-4.

SERVENTI, F., RAMAZZINA, I., LAMBERTO, I., PUGGIONI, V., GATTI, R. & PERCUDANI, R. 2010. Chemical basis of nitrogen recovery through the ureide pathway: formation and hydrolysis of S-ureidoglycine in plants and bacteria. *ACS Chem Biol,* 5**,** 203-14.

SETO, H., JINNAI, Y., HIRATSUKA, T., FUKAWA, M., FURIHATA, K., ITOH, N. & DAIRI, T. 2008. Studies on a new biosynthetic pathway for menaquinone. *J Am Chem Soc,* 130**,** 5614-5.

SHLOMI, T., CABILI, M. N., HERRGARD, M. J., PALSSON, B. O. & RUPPIN, E. 2008. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology,* 26**,** 1003-10.

SHOJI, S., DAMBACHER, C. M., SHAJANI, Z., WILLIAMSON, J. R. & SCHULTZ, P. G. 2011. Systematic chromosomal deletion of bacterial ribosomal protein genes. *J Mol Biol,* 413**,** 751-61.

SHOVAL, O., SHEFTEL, H., SHINAR, G., HART, Y., RAMOTE, O., MAYO, A., DEKEL, E., KAVANAGH, K. & ALON, U. 2012. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science,* 336**,** 1157-60.

SMITH, E. A. & MACFARLANE, G. T. 1996. Enumeration of human colonic bacteria producing phenolic and indolic compounds: effects of pH, carbohydrate availability and retention time on dissimilatory aromatic amino acid metabolism. *J Appl Bacteriol,* 81**,** 288-302.

SONG, Y., LEE, B. R., CHO, S., CHO, Y. B., KIM, S. W., KANG, T. J., KIM, S. C. & CHO, B. K. 2015. Determination of single nucleotide variants in Escherichia coli DH5alpha by using short-read sequencing. *FEMS Microbiol Lett,* 362.

SPENCER, J. B., STOLOWICH, N. J., ROESSNER, C. A. & SCOTT, A. I. 1993. The Escherichia coli cysG gene encodes the multifunctional protein, siroheme synthase. *FEBS Lett,* 335**,** 57-60.

SPRENGER, G. A. 1995. Genetics of pentose-phosphate pathway enzymes of Escherichia coli K-12. *Arch Microbiol,* 164**,** 324-30.

STALHEIM, O. H. & WILSON, J. B. 1964. Cultivation of Leptospirae. I. Nutrition of Leptospira Canicola. *J Bacteriol,* 88**,** 48-54.

STANECK, J. L., HENNEBERRY, R. C. & COX, C. D. 1973. Growth requirements of pathogenic Leptospira. *Infect Immun,* 7**,** 886-97.

STERN, N., SHENBERG, E. & TIETZ, A. 1969. Studies on the metabolism of fatty acids in Leptospira: the biosynthesis of delta 9- and delta 11-monounsaturated acids. *Eur J Biochem,* 8**,** 101-8.

STOLYAR, S., VAN DIEN, S., HILLESLAND, K. L., PINEL, N., LIE, T. J., LEIGH, J. A. & STAHL, D. A. 2007. Metabolic modeling of a mutualistic microbial community. *Molecular systems biology,* 3**,** 92.

SUBBARAYAN, P. R. & SARKAR, M. 2004. A comparative study of variation in codon 33 of the rpoS gene in Escherichia coli K12 stocks: implications for the synthesis of sigma(s). *Mol Genet Genomics,* 270**,** 533-8.

SUN, Y., FLEMING, R. M., THIELE, I. & SAUNDERS, M. A. 2013. Robust flux balance analysis of multiscale biochemical reaction networks. *BMC bioinformatics,* 14**,** 240.

SUNG, J. M. & LINDSAY, J. A. 2007. Staphylococcus aureus strains that are hypersusceptible to resistance gene transfer from enterococci. *Antimicrob Agents Chemother,* 51**,** 2189-91.

SUTCLIFFE, I. C. & SHAW, N. 1991. Atypical lipoteichoic acids of gram-positive bacteria. *J Bacteriol,* 173**,** 7065-9.

SWAINSTON, N., SMALLBONE, K., MENDES, P., KELL, D. & PATON, N. 2011. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of integrative bioinformatics,* 8**,** 186.

SZAPPANOS, B., KOVACS, K., SZAMECZ, B., HONTI, F., COSTANZO, M., BARYSHNIKOVA, A., GELIUS-DIETRICH, G., LERCHER, M. J., JELASITY, M., MYERS, C. L., ANDREWS, B. J., BOONE, C., OLIVER, S. G., PAL, C. & PAPP, B. 2011. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet,* 43**,** 656-62.

TATTEVIN, P., DIEP, B., JULA, M. & PERDREAU-REMINGTON, F. 2009. Methicillin-resistant Staphylococcus aureus USA300 clone in long-term care facility. *Emerg*

Infect Dis [serial on the Internet]. [Online]. Available: Available from http://wwwnc.cdc.gov/eid/article/15/6/08-0195.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. & NATALE, D. A. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4**,** 41.

TAYLOR, R. G., WALKER, D. C. & MCINNES, R. R. 1993. E. coli host strains significantly affect the quality of small scale plasmid DNA preparations used for sequencing. *Nucleic Acids Res,* 21**,** 1677-8.

TENOVER, F. C. & GOERING, R. V. 2009. Methicillin-resistant Staphylococcus aureus strain USA300: origin and epidemiology. *J Antimicrob Chemother,* 64**,** 441-6.

TEPPER, N. & SHLOMI, T. 2011. Computational design of auxotrophy-dependent microbial biosensors for combinatorial metabolic engineering experiments. *PloS one,* 6**,** e16274.

TETTELIN, H., RILEY, D., CATTUTO, C. & MEDINI, D. 2008. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol,* 11**,** 472-7.

THEODORE, T. S. & PANOS, C. 1973. Protein and fatty acid composition of mesosomal vesicles and plasma membranes of Staphylococcus aureus. *J Bacteriol,* 116**,** 571-6.

THIELE, I., FLEMING, R. M., BORDBAR, A., SCHELLENBERGER, J. & PALSSON, B. O. 2010. Functional characterization of alternate optimal solutions of Escherichia coli's transcriptional and translational machinery. *Biophysical journal,* 98**,** 2072-81.

THIELE, I., FLEMING, R. M., QUE, R., BORDBAR, A., DIEP, D. & PALSSON, B. O. 2012. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. *PloS one,* 7**,** e45635.

THIELE, I., HYDUKE, D. R., STEEB, B., FANKAM, G., ALLEN, D. K., BAZZANI, S., CHARUSANTI, P., CHEN, F. C., FLEMING, R. M., HSIUNG, C. A., DE KEERSMAECKER, S. C., LIAO, Y. C., MARCHAL, K., MO, M. L., OZDEMIR, E., RAGHUNATHAN, A., REED, J. L., SHIN, S. I., SIGURBJORNSDOTTIR, S., STEINMANN, J., SUDARSAN, S., SWAINSTON, N., THIJS, I. M., ZENGLER, K., PALSSON, B. O., ADKINS, J. N. & BUMANN, D. 2011. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst Biol,* 5**,** 8.

THIELE, I., JAMSHIDI, N., FLEMING, R. & PALSSON, B. 2009a. 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.

THIELE, I., JAMSHIDI, N., FLEMING, R. M. & PALSSON, B. O. 2009b. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS computational biology,* 5**,** e1000312.

THIELE, I. & PALSSON, B. O. 2010a. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols,* 5**,** 93-121.

THIELE, I. & PALSSON, B. O. 2010b. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc,* 5**,** 93-121.

THIELE, I., SWAINSTON, N., FLEMING, R. M., HOPPE, A., SAHOO, S., AURICH, M. K., HARALDSDOTTIR, H., MO, M. L., ROLFSSON, O., STOBBE, M. D., THORLEIFSSON, S. G., AGREN, R., BOLLING, C., BORDEL, S., CHAVALI, A. K., DOBSON, P., DUNN, W. B., ENDLER, L., HALA, D., HUCKA, M., HULL, D., JAMESON, D., JAMSHIDI, N., JONSSON, J. J., JUTY, N., KEATING, S., NOOKAEW, I., LE NOVERE, N., MALYS, N., MAZEIN, A., PAPIN, J. A., PRICE, N. D., SELKOV, E., SR., SIGURDSSON, M. I., SIMEONIDIS, E., SONNENSCHEIN, N., SMALLBONE, K., SOROKIN, A., VAN BEEK, J. H., WEICHART, D., GORYANIN, I., NIELSEN, J., WESTERHOFF, H. V., KELL, D. B., MENDES, P. & PALSSON, B. O. 2013a. A community-driven global reconstruction of human metabolism. *Nat Biotechnol,* 31**,** 419-25.

THIELE, I., SWAINSTON, N., FLEMING, R. M., HOPPE, A., SAHOO, S., AURICH, M. K., HARALDSDOTTIR, H., MO, M. L., ROLFSSON, O., STOBBE, M. D., THORLEIFSSON, S. G., AGREN, R., BOLLING, C., BORDEL, S., CHAVALI, A. K., DOBSON, P., DUNN, W. B., ENDLER, L., HALA, D., HUCKA, M., HULL, D., JAMESON, D., JAMSHIDI, N., JONSSON, J. J., JUTY, N., KEATING, S., NOOKAEW, I., LE NOVERE, N., MALYS, N., MAZEIN, A., PAPIN, J. A., PRICE, N. D., SELKOV, E., SR., SIGURDSSON, M. I., SIMEONIDIS, E., SONNENSCHEIN, N., SMALLBONE, K., SOROKIN, A., VAN BEEK, J. H., WEICHART, D., GORYANIN, I., NIELSEN, J., WESTERHOFF, H. V., KELL, D. B., MENDES, P. & PALSSON, B. O. 2013b. A community-driven global reconstruction of human metabolism. *Nature biotechnology,* 31**,** 419-25.

THORLEIFSSON, S. G. & THIELE, I. 2011. rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics,* 27**,** 2009-10.

TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C., BOUVET, O., CALTEAU, A., CHIAPELLO, H., CLERMONT, O., CRUVEILLER, S., DANCHIN, A., DIARD, M., DOSSAT, C., KAROUI, M. E., FRAPY, E., GARRY, L., GHIGO, J. M., GILLES, A. M., JOHNSON, J., LE BOUGUENEC, C., LESCAT, M., MANGENOT, S., MARTINEZ-JEHANNE, V., MATIC, I., NASSIF, X., OZTAS, S., PETIT, M. A., PICHON, C., ROUY, Z., RUF, C. S., SCHNEIDER, D., TOURRET, J., VACHERIE, B., VALLENET, D., MEDIGUE, C., ROCHA, E. P. & DENAMUR, E. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet,* 5**,** e1000344.

TSUNEKAWA, H., AZUMA, S., OKABE, M., OKAMOTO, R. & AIBA, S. 1992. Acquisition of a sucrose utilization system in Escherichia coli K-12 derivatives and its application to industry. *Appl Environ Microbiol,* 58**,** 2081-8.

TYNECKA, Z., SZCZESNIAK, Z., MALM, A. & LOS, R. 1999. Energy conservation in aerobically grown Staphylococcus aureus. *Res Microbiol,* 150**,** 555-66.

VAREMO, L., NIELSEN, J. & NOOKAEW, I. 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res,* 41**,** 4378-91.

VARMA, A., BOESCH, B. W. & PALSSON, B. O. 1993. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Appl Environ Microbiol,* 59**,** 2465-73.

VIEIRA, G., SABARLY, V., BOURGUIGNON, P. Y., DUROT, M., LE FEVRE, F., MORNICO, D., VALLENET, D., BOUVET, O., DENAMUR, E., SCHACHTER, V. & MEDIGUE, C. 2011. Core and panmetabolism in Escherichia coli. *J Bacteriol,* 193**,** 1461-72.

VIJAYENDRAN, C., POLEN, T., WENDISCH, V. F., FRIEHS, K., NIEHAUS, K. & FLASCHEL, E. 2007. The plasticity of global proteome and genome expression analyzed in closely related W3110 and MG1655 strains of a well-studied model organism, Escherichia coli-K12. *J Biotechnol,* 128**,** 747-61.

VINNIKOV, A. I. 1988. [ATP synthesis in Staphylococcus aureus cells during induction of membrane potentials and proton gradient]. *Biokhimiia,* 53**,** 853-5.

VITKIN, E. & SHLOMI, T. 2012. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biol,* 13**,** R111.

VOLLMER, W. 2012. Bacterial growth does require peptidoglycan hydrolases. *Mol Microbiol,* 86**,** 1031-5.

WAACK, S., KELLER, O., ASPER, R., BRODAG, T., DAMM, C., FRICKE, W. F., SUROVCIK, K., MEINICKE, P. & MERKL, R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics,* 7**,** 142.

WANNET, W. J., SPALBURG, E., HECK, M. E., PLUISTER, G. N., TIEMERSMA, E., WILLEMS, R. J., HUIJSDENS, X. W., DE NEELING, A. J. & ETIENNE, J. 2005. Emergence of virulent methicillin-resistant Staphylococcus aureus strains carrying Panton-Valentine leucocidin genes in The Netherlands. *J Clin Microbiol,* 43**,** 3341-5.

WEIGEL, L. M., CLEWELL, D. B., GILL, S. R., CLARK, N. C., MCDOUGAL, L. K., FLANNAGAN, S. E., KOLONAY, J. F., SHETTY, J., KILLGORE, G. E. &

TENOVER, F. C. 2003. Genetic analysis of a high-level vancomycin-resistant isolate of Staphylococcus aureus. *Science,* 302**,** 1569-71.

WHITE, D. C. & FRERMAN, F. E. 1967. Extraction, characterization, and cellular localization of the lipids of Staphylococcus aureus. *J Bacteriol,* 94**,** 1854-67.

WHO 2003. Human Leptospirosis: Guidance for Diagnosis, Surveillance and Control. *In:* ORGANIZATION, W. H. (ed.). Geneva, Switzerland.

WHO 2014. Antimicrobial resistance: global report on surveillance 2014. World health organization.

WILKINSON, B. J., CROSSLEY, K. & ARCHER, G. 1997. *The staphylococci in human disease*, Churchill Livingston.

WILLIAMS, T. C., POOLMAN, M. G., HOWDEN, A. J., SCHWARZLANDER, M., FELL, D. A., RATCLIFFE, R. G. & SWEETLOVE, L. J. 2010. A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant physiology,* 154**,** 311-23.

WINTERMUTE, E. H. & SILVER, P. A. 2010. Emergent cooperation in microbial metabolism. *Molecular systems biology,* 6**,** 407.

WU, D., HUGENHOLTZ, P., MAVROMATIS, K., PUKALL, R., DALIN, E., IVANOVA, N. N., KUNIN, V., GOODWIN, L., WU, M., TINDALL, B. J., HOOPER, S. D., PATI, A., LYKIDIS, A., SPRING, S., ANDERSON, I. J., D'HAESELEER, P., ZEMLA, A., SINGER, M., LAPIDUS, A., NOLAN, M., COPELAND, A., HAN, C., CHEN, F., CHENG, J. F., LUCAS, S., KERFELD, C., LANG, E., GRONOW, S., CHAIN, P., BRUCE, D., RUBIN, E. M., KYRPIDES, N. C., KLENK, H. P. & EISEN, J. A. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature,* 462**,** 1056-60.

YAMAMOTO, N., NAKAHIGASHI, K., NAKAMICHI, T., YOSHINO, M., TAKAI, Y., TOUDA, Y., FURUBAYASHI, A., KINJYO, S., DOSE, H., HASEGAWA, M., DATSENKO, K. A., NAKAYASHIKI, T., TOMITA, M., WANNER, B. L. & MORI, H. 2009. Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Molecular systems biology,* 5**,** 335.

YANG, C. W., WU, M. S. & PAN, M. J. 2001. Leptospirosis renal disease. *Nephrol Dial Transplant,* 16 Suppl 5**,** 73-7.

YE, Y. & GODZIK, A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics,* 19 Suppl 2**,** ii246-55.

YIM, H., HASELBECK, R., NIU, W., PUJOL-BAXLEY, C., BURGARD, A., BOLDT, J., KHANDURINA, J., TRAWICK, J. D., OSTERHOUT, R. E., STEPHEN, R., ESTADILLA, J., TEISAN, S., SCHREYER, H. B., ANDRAE, S., YANG, T. H., LEE, S. Y., BURK, M. J. & VAN DIEN, S. 2011. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol,* 7**,** 445-52.

YOON, S. H., HAN, M. J., JEONG, H., LEE, C. H., XIA, X. X., LEE, D. H., SHIM, J. H., LEE, S. Y., OH, T. K. & KIM, J. F. 2012. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biol,* 13**,** R37.

YOSHIDA, A. & DAVE, V. 1975. Inhibition of NADP-dependent dehydrogenases by modified products of NADPH. *Arch Biochem Biophys,* 169**,** 298-303.

YUM, D. Y., LEE, B. Y. & PAN, J. G. 1999. Identification of the yqhE and yafB genes encoding two 2, 5-diketo-D-gluconate reductases in Escherichia coli. *Appl Environ Microbiol,* 65**,** 3341-6.

YUTIN, N., PUIGBO, P., KOONIN, E. V. & WOLF, Y. I. 2012. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One,* 7**,** e36972.

ZAGAGLIA, C., CASALINO, M., COLONNA, B., CONTI, C., CALCONI, A. & NICOLETTI, M. 1991. Virulence plasmids of enteroinvasive Escherichia coli and Shigella flexneri integrate into a specific site on the host chromosome: integration greatly reduces expression of plasmid-carried virulence genes. *Infect Immun,* 59**,** 792-9.

ZAMBONI, N., FENDT, S. M., RUHL, M. & SAUER, U. 2009. (13)C-based metabolic flux analysis. *Nature protocols,* 4**,** 878-92.

ZHANG, Q., ZHANG, Y., ZHONG, Y., MA, J., PENG, N., CAO, X., YANG, C., ZENG, R., GUO, X. & ZHAO, G. 2011. Leptospira interrogans encodes an ROK family glucokinase involved in a cryptic glucose utilization pathway. *Acta Biochim Biophys Sin (Shanghai),* 43**,** 618-29.

ZHANG, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,* 9**,** 40.

ZHANG, Y., THIELE, I., WEEKES, D., LI, Z., JAROSZEWSKI, L., GINALSKI, K., DEACON, A. M., WOOLEY, J., LESLEY, S. A., WILSON, I. A., PALSSON, B., OSTERMAN, A. & GODZIK, A. 2009. Three-dimensional structural view of the central metabolic network of Thermotoga maritima. *Science,* 325**,** 1544-9.

ZHAO, S., KUMAR, R., SAKAI, A., VETTING, M. W., WOOD, B. M., BROWN, S., BONANNO, J. B., HILLERICH, B. S., SEIDEL, R. D., BABBITT, P. C., ALMO, S. C., SWEEDLER, J. V., GERLT, J. A., CRONAN, J. E. & JACOBSON, M. P. 2013. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*.

ZHU, Y., WEISS, E. C., OTTO, M., FEY, P. D., SMELTZER, M. S. & SOMERVILLE, G. A. 2007. Staphylococcus aureus biofilm metabolism and the influence of arginine on polysaccharide intercellular adhesin synthesis, biofilm formation, and pathogenesis. *Infect Immun,* 75**,** 4219-26.

ZHUANG, K., IZALLALEN, M., MOUSER, P., RICHTER, H., RISSO, C., MAHADEVAN, R. & LOVLEY, D. R. 2011a. Genome-scale dynamic modeling of the competition

between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME journal,* 5**,** 305-16.

ZHUANG, K., VEMURI, G. N. & MAHADEVAN, R. 2011b. Economics of membrane occupancy and respiro-fermentation. *Molecular systems biology,* 7**,** 500.