UNIVERSITY OF CALIFORNIA, SAN DIEGO

Systems level decomposition of adaptation to metabolic perturbation

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

in

Bioengineering

by

Douglas McCloskey

Committee in charge:

Bernhard Ø. Palsson, Chair
Pedro Cabrales
Adam M. Feist
Christian Metallo
Robert K. Naviaux
John Watson

2017

The Dissertation of Douglas McCloskey is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reformatted reprint of “Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli.” McCloskey D, Palsson BO, Feist AM. Mol Syst Biol. 2013;9:661. doi: 10.1038/msb.2013.18.. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 3, in full, is a reformatted reprint of “A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent.” McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, Feist AM. Biotechnol Bioeng. 2014 Apr;111(4):803-15. doi: 10.1002/bit.25133. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 4, in full, is a reformatted reprint of “Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media.” McCloskey, D, Utrilla, J, Naviaux, RK, Palsson BO, Feist AM. Metabolomics (2015) 11: 198. doi:10.1007/s11306-014-0686-2. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 5, in full, is a reformatted reprint of “A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites.” McCloskey, D, Gangoiti, JA, Palsson BO, Feist AM. Metabolomics (2015) 11: 1338.

doi:10.1007/s11306-015-0790-y. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 6, in full, is a reformatted reprint of “MID Max: LC–MS/MS Method for Measuring the Precursor and Product Mass Isotopomer Distributions of Metabolic Intermediates and Cofactors for Metabolic Flux Analysis Applications.” McCloskey D, Young JD, Xu S, Palsson BO, Feist AM. Anal Chem. 2016 Jan 19;88(2):1362-70. doi: 10.1021/acs.analchem.5b03887. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 7, in full, is a reformatted reprint of “Modeling Method for Increased Precision and Scope of Directly Measurable Fluxes at a Genome-Scale.” McCloskey D, Young JD, Xu S, Palsson BO, Feist AM. Anal Chem. 2016 Apr 5;88(7):3844-52. doi: 10.1021/acs.analchem.5b04914. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 8, in full, is a reformatted submission of “Laboratory Evolution of Gene Knockout Strains Reveals Fundamental Principles of Adaptation.” McCloskey D., Xu S., Sandberg T.E., Brunk E., Hefner Y., Szubin R., Feist A.M., and Palsson BO. Cell. The dissertation/thesis author was the primary investigator and author of the paper.

VITA

2010 Bachelors of Science, University of California, Irvine

2017 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Nguyen, D., McCloskey (Taylor), D., et al. Better shrinkage than Shrinky-Dinks. Lab Chip 10, 1623-1626 (2010).

McCloskey (Taylor), D., Dyer, D., Lew, V. & Khine, M. Shrink film patterning by craft cutter: complete plastic chips with high resolution/high-aspect ratio channel. Lab Chip 10, 2472-2475 (2010).

McCloskey D, Palsson BØ, Feist AM. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol. 9:661.

Akizu, N., Cantagrel, V., Schroth, J., Cai, N., Vaux, K., McCloskey, D., Gleeson, J. G. (2013). AMPD2 Regulates GTP Synthesis and is Mutated in a Potentially-Treatable Neurodegenerative Brainstem Disorder. Cell.

McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, Feist AM. (2013). A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. Biotechnol Bioeng.

Lewis CA, Parker SJ, Fiske BP, McCloskey D, Gui DY, Green CR, Vokes NI, Feist AM, Heiden MGVander, Metallo CM. (2014). Tracing Compartmentalized NADPH Metabolism in the Cytosol and Mitochondria of Mammalian Cells. Mol Cell.

McCloskey D, Utrilla J, Naviaux RK, Palsson BO, Feist AM. (2014). Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. Metabolomics.

Olavarria K, De Ingeniis J, Zielinski DC, Fuentealba M, Muñoz R, McCloskey D, Feist AM, Cabrera R. (2014). The Metabolic Impact of a NADH-producing Glucose-6-phosphate Dehydrogenase in Escherichia coli. Microbiology.

McCloskey D, Gangoiti JA, Palsson BO, Feist AM. (2015). A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites. Metabolomics.

Bordbar A, McCloskey D, Zielinski D, Sonnenschein N, Jamshidi N, Palsson BO (2015). Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. Cell Systems.

McCloskey D, Young JD, Xu S, Palsson BO, and Feist AM (2016). MID Max: LC–MS/MS Method for Measuring the Precursor and Product Mass Isotopomer Distributions of Metabolic Intermediates and Cofactors for Metabolic Flux Analysis Applications. Anal Chem.

McCloskey D, Young JD, Xu S, Palsson BØ, Feist AM. (2016). A modeling method for increased precision and scope of directly measurable fluxes at a genome-scale. Anal Chem.

Utrilla J, O'Brien E, Chen K, McCloskey D, Cheung J, Wang H, Armenta-Medina D, Feist AM, Palsson BØ (2016). Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. Cell Systems. DOI: 10.1016/j.cels.2016.04.003

Brunk E, George KW, Alonso-Gutierrez J, Thompson M, Baidoo E, Wang G, Petzold CJ, McCloskey D, Monk J, Yang L, O'Brien EJ, Batth TS, Martin HGarcia, Feist A, Adams PD, Keasling JD, Palsson BO, Lee TSoon (2016). Characterizing Strain Variation in Engineered E. coli Using a Multi-Omics-Based Workflow. Cell Syst.

ABSTRACT OF THE DISSERTATION

Systems level decomposition of adaptation to metabolic perturbation

by

Douglas McCloskey

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2017

Professor Bernhard O. Palsson, Chair

Much progress has been made in establishing the causality of mutations that occur during adaptive laboratory evolution (ALE) on organism physiology. In contrast, little progress has been made in detailing the mechanisms and overarching principles of evolution that govern the adaptive process. This dissertation describes the following three aims that sought to address the technical and scientific challenges required to better understand adaptive

evolution in micro-organisms. First, an improved method to quantify intracellular metabolites from sampling and extraction to acquisition and quantitation was developed and validated. Second, an improved method to measure intracellular fluxes from sampling and extraction to acquisition and modeling was developed and validated. And finally, aims 1 and 2 were applied in order to uncover system level principles and mechanistic level network functions that were required to adapt to major metabolic perturbation.

CHAPTER 1:

Introduction

How do organisms adapt to metabolic perturbation? Using *E. coli* as a model organism, this dissertation documents the process by which that question was addressed. This dissertation also describes the series of methodological, analytical, scientific contributions that were made while addressing that question.

A review of the state of biochemical modeling applications using *E. coli* at the time of starting the doctoral degree was first conducted (Chapter 2). The review describes the 6 categories of biochemical modelling uses. The review also identifies several gaps that hinder our ability to accurately predict biological function. These gaps pertain to the scope of biological processes that are not incorporated into the model (e.g., the transcription regulatory network, expression, DNA synthesis, etc.,) and regulation of biological interactions (e.g., small molecule allosteric regulation of enzymes, small molecule activation of transcription factors, etc.). Two of the most important data types required to fill these gaps were metabolomics and fluxomics.

Metabolomics is the systematic study of the unique chemical fingerprints (i.e., small molecule metabolite profiles) that specific cellular processes leave behind[1]. Metabolomics provides a “snapshop of cellular physiology” that can be used to identify biomarkers and quantitatively model biochemical reactions[2].
In order to identify the most important metabolites to include in that “snapshot”, a modeling method to identify the most highly utilized metabolites in a genome-

scale model of metabolism from which a targeted assay can be constructed was described (Chapter 3). A first iteration of that assay was then constructed. The derived data was then integrated with thermodynamic analysis to study the metabolome of anaerobic and aerobic wild-type *E. coli*.

Several limitations to the sampling and extraction method and metabolite acquisition method were identified during the work described in chapter 3 that were addressed in (chapters 4 and 5). To this end, an improved sampling and extraction method for intracellular metabolomics was developed and validated (chapter 4). The method was shown to provide fast and rapid sampling from a diverse set of culturing conditions including both aerobic and anaerobic culturing vesicles. In addition, a reverse phase ion-pairing (RIP) liquid chromatography (LC) mass spectrometry[2] (MS/MS) method for improved quantitation of intracellular metabolites was developed and validated (chapter 5). The method was shown to provide a reduction in carryover, superior separation of many biological isomers, and improved accuracy and confidence in metabolite identification compared to previously published methods.

Fluxomics describes the approaches to calculate *in vivo* reaction rates (i.e., fluxes) of biological entities[3]. The fluxome, or conglomerate of all enzymatic reaction rates of a biological entity, is said to describe the cellular phenotype or functional state of the cell. The majority of fluxomics acquisition methods are limited to the steady-state measurement of proteogenic amino acids. Those methods that do measure the isotope labeling patterns of intracellular metabolites for non-steady-state and dynamic experiments are often limited in

the number of metabolites and metabolite fragments that can be measured. In addition, the majority of fluxomics modelling methods to incorporate isotope labeling patterns with biochemical networks are limited in scope to simplified models of central metabolism. These methods are not able to calculate reaction fluxes outside of central metabolism that have important consequences on cellular physiology.

These limitations to the field of fluxomics were addressed (chapters 6 and 7). A RIP-LC-MS/MS acquisition method for improved coverage of intracellular mass isotopomer distributions (MIDs) for metabolic flux analysis (MFA) applications was developed and validated. The method was shown to expand the number of MIDs that can be measure in a single run, and was shown to allow for the measurement of MIDs of cofactors and other peripheral pathway metabolites. In addition, a genome-scale MFA modeling method for expanded coverage of directly measurable fluxes was developed and validated. The method incorporated the improved measurement abilities described previously in combination with a genome-scale MFA model that was shown to calculate an increased number and scope of reaction fluxes without a loss in flux precision.

Adaptive laboratory evolution (ALE) is an experimental method that introduces a selection pressure (e.g., growth rate selection) in a controlled environmental setting[4]. Using ALE, organisms can be perturbed from their evolutionary optimized homeostatic states, and their re-adjustments studied during the course of adaptation to reveal novel and non-intuitive component functions and interactions. The metabolomics and fluxomics methods that were

developed and validated (chapters 4-7) along with a bioinformatics software pipeline were applied to uncover the systems level and mechanistic level changes that occur following metabolic perturbation and ALE (chapter 8).

References:

1. Johnson, C.H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* **17**, 451-459 (2016).

2. Jamshidi, N. & Palsson, B.Ø. Mass Action Stoichiometric Simulation Models: Incorporating Kinetics and Regulation into Stoichiometric Models. *Biophysical Journal* **98**, 175-185 (2010).

3. Feng, X. Page, L.,Rubens, J.,Chircus, L.,Colletti, P., Pakrasi, H.B., and Tang Y.J. Bridging the Gap between Fluxomics and Industrial Biotechnology. *Journal of Biomedicine and Biotechnology* **2010** (2010).

4. Tenaillon, O. Barrick J.E., Ribeck N., Deatherage D.E., Blanchard J.L., Dasgupta A., Wu G.C., Wielgoss S., Cruveiller S., Médigue C., Schneider D., Lenski R.E. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165-170 (2016).

CHAPTER 2:

Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*

Abstract

The genome-scale model (GEM) of metabolism in the bacterium *Escherichia coli* K-12 has been in development for over a decade and is now in wide use. GEM-enabled studies of E. coli have been primarily focused on six applications: 1) metabolic engineering, 2) model-driven discovery, 3) prediction of cellular phenotypes, 4) analysis of biological network properties, 5) studies of evolutionary processes, and 6) models of interspecies interactions. In this Review, we provide an overview of these applications, along with a critical assessment of their successes and limitations and a perspective on likely future developments in the field. Taken together, the studies performed over the past decade have established a genome-scale mechanistic understanding of genotype-phenotype relationships in *E. coli* metabolism that is forming the basis for similar efforts for other microbial species. Future challenges include the expansion of GEMs by integrating additional cellular processes beyond metabolism and the development of computational methods able to handle such large-scale network models with sufficient power and accuracy.

Introduction

Whole genome sequencing along with decades worth of detailed biochemical and enzymatic data (e.g., bibliomic data) on microbial metabolism has led to the reconstruction of metabolic networks at the genome-scale (so called, GENREs or genome-scale reconstructions) [1-4]. Integrating this information in a structured fashion has enabled its translation into computational models that can be used to calculate metabolic phenotypes [5-7]. In addition, other omics data types that have been generated can be interpreted in the context of a reconstruction and computational model to analyze cellular functions under specific conditions. Taken together, this information becomes a *de facto* knowledge base. Genome-scale models (GEMs) are a structured format of such a knowledge base that can be used to perform computational and quantitative queries to answer various questions about the capabilities of organisms and their likely phenotypic states [8, 9].

Escherichia coli is one of the most important model organisms in biology and its metabolic GEM has aided the development of microbial systems biology. The history of the metabolic network reconstruction process for *E. coli* and the formulation and testing of its metabolic GEM now spans over a decade (Figure 1). One of the earliest studies to systematically analyze *E. coli* utilized a simplified constraint-based model of acetate overflow [10]. Subsequent pre-genome-scale studies expanded upon the constraint-based approach to include reactions involved in central carbohydrate metabolism, amino acid synthesis, and nucleotide synthesis to evaluate the biocatalyst production potential of *E. coli* [11,

[12] using flux balance analysis (FBA) [13]. The ability of FBA, in particular, and constraint-based modeling, in general, to quantitatively describe the metabolic physiology of *E. coli* observed experimentally [14] arguable solidified the value of systems level analysis in understanding microbial metabolism. The sequencing of the *E. coli* genome [15], the advent of the lambda-red system for efficient genome manipulation [16], and information readily available on annotated content of *E. coli* in databases and detailed biochemical reviews led to a steady increase in content of the *E. coli* GEM in the genomic era. Reconstruction efforts in the 2000s built off of successive versions, each adding new subsystems (e.g., fatty acid, alternate carbon metabolism, and cell wall synthesis, respectively) as the reconstructions strived to incorporate all of the existing content in literature and newly appearing data. Analysis of the rate at which new content was added to the latest metabolic GEM [17] indicates that mostly newly characterized content is now left to include in the reconstruction. Future expansion for the metabolic GEM is likely to come from characterizing promiscuity of known enzymes and the addition of protein synthesis will open the door for more detailed examination of other cellular processes and integration with other omics data sets.

In over a decade of model-driven development of systems biology for *E. coli*, over 200 peer-reviewed studies have appeared, as summarized in Figures 2 and Supplementary Figure 1, and documented in greater detail in Supplementary Table 1. This review aims to determine the benefits and drawbacks of the uses of the *E. coli* GEM to date, what has been accomplished, what has been missed, and what is likely to lie ahead for this field in its next decade of development.

Categories of Uses of GEMs

The *E. coli* GEM has been applied to answer different biological questions (Fig. 3), most frequently in the categories of: 1) metabolic engineering, 2) model-driven discovery, 3) prediction of cellular phenotypes, 4) analysis of biological network properties, 5) studies of evolutionary processes, and 6) models of interspecies interactions. In a previous review, we described and categorized the uses of the *E. coli* GEM appearing in 64 papers published prior to 2007 [18]. Similarly, Oberhardt et al. (Oberhardt et al, 2009) have reviewed applications enabled by GEMs for organisms other than *E. coli* through 2009, and specific reviews on GEM-enabled studies in plants [19] and *Saccharomyces cerevisiae* [20] have recently appeared. However, the *E. coli* GEM remains the oldest and arguably most extensively utilized GEM, and given the extensive uses of the *E. coli* GEM that have appeared since 2007, we now have a sufficient amount of studies to critically assess what GEMs can and cannot do by focusing on the *E. coli* GEM as a subject. For example, metabolic engineering and model-driven discovery are categories of uses in *E. coli* that have matured over the years into workflows that can be continuously repeated to tackle a diverse set of biological questions. In particular, strain design has matured from academic to industrial thanks to the advent of sustainable processes that can be applied from one target compound to another. The development of these iterative workflows in systems biology is a theme that will be highlighted and discussed in greater detail as they appear in each category. In another example, recent studies modeling interspecies interactions signal an expansion of the field from single cell models

to ecosystem level models. We have also highlighted a collection of noteworthy studies (Supplementary Figure 2) from the pool of uses that we feel demonstrate how a GEM can be well utilized to deduce biological complexities.

A detailed overview of the successes and limitations of the *E. coli* GEM implementation for the categories outlined above have been summarized in Table 1 and will be presented in detail in each section below. At the conclusion of each section, we will then assess how the limitations of GEMs can be overcome with further development and refinement. We also highlight systems biology workflows made possible by GEMs. Furthermore, we will attempt to identify the future challenges that need to be overcome to bring us closer to the ultimate goal of establishing a comprehensive and multi-scale mechanistic understanding of the genotype-phenotype relationship of microbial metabolism. It should be stressed, that although this review is focused on the *E. coli* GEM, the findings documented here can be readily extended to metabolic modeling at the genome-scale in general.

Category of uses of GEMs:

1. Metabolic engineering

Current industrial processes rely upon non-renewable resources that cannot sustain the growing world population indefinitely. The development of biosustainable processes that can convert renewable resources into commodity items is therefore of paramount socio-economic importance. Bacteria have recently emerged as a means by which to achieve bio sustainability [21]. Through metabolic engineering, the native biochemical pathways of bacteria can be manipulated and optimized to more efficiently produce industrial and therapeutically relevant compounds. The *E. coli* GEM has guided metabolic engineers towards the production of an assortment of compounds including organic acids, amino acids, and alcohols to name a few (see Supplementary Table 3 for a comprehensive list).

Contrary to random mutagenesis and screening, rational strain design uses the GEM to predict cellular phenotypes from a systems level using genomic, stoichiometric, kinetic, and regulatory knowledge to identify engineering strategies, which can then be implemented *in vivo*. These strategies include gene deletions [22], gene over and under expression [23], mapping high throughput data onto the network reconstruction to identify bottlenecks or competing pathways [24], and more recently, integration of non-native pathways into standard microbial production hosts for production of compounds that are either not natively found in, or only synthesized in minute concentrations by the host [25-27]. More advanced methods have even allowed for the identification of strategies

that couple bacterial growth to target product overproduction [28-30]. A so-called "growth-coupled" strain leads to a more robust strain that is less likely to lose the genetically engineered genotype, or be outcompeted by alternate bacterial phenotypes in a bioprocess environment. A cadre of algorithms, first appearing in 2003 [30], with increasing biological detail [29] or alternative optimization methods [28] have been developed (Supplementary Table 2) and extensively reviewed [31, 32].

Engineering strategies found using model-driven analysis can often be non-intuitive and highlight some of the most interesting recent findings using GEMs. For instance, researchers used the GEM to not only determine a gene which needed to be upregulated, but were able to tune the expression level of this gene after subsequent GEM analysis of a deleterious overexpression event[24]. In another GEM-enabled study, the highest flavanone production was predicted and experimentally determined by strategically knocking out genes to not only increase the production of the redox carrier (NADPH) to drive the heterologous flavanone catalyst, but also to maintain the optimal redox potential of the cell (i.e., the ratio of NADPH to $NADP^+$) [23, 33]. For a final example, researchers improved the production potential of the non-native metabolite 1,4-butanediol in *E. coli* by over three orders of magnitude [25]. The researchers were able to rewire the host cell to produce the compound via native and non-native pathways by ensuring that the production of 1,4-butanediol was the only means by which the host cell could maintain redox balance and grow anaerobically [25]. These successes highlight the need to analyze genetic alterations at the systems level where one can not only predict the activation of pathways that compensate

for lost functionalities following gene deletions, but also predict engineering strategies that couple cellular goals to target compound overproduction.

Engineering strategies derived from the *E. coli* GEM have also led to non-viable and suboptimal phenotypes. Even in the above studies [23, 25, 33], the authors had to carefully select which predicted knockout designs were constructed *in vivo*, due to, for example, known isozymes that result in non-viability when deleted simultaneously. Hence, a strong understanding of metabolic biochemistry is a prerequisite for successful strain design. Many potential engineering strategies cannot be addressed with the current generation of GEMs as they do not account for translational regulation and detailed enzyme kinetics. For example, strains generated using random knockouts via transposon libraries and screening for lycopene overproduction identified gene deletion targets in regulatory elements [34] that cannot be predicted by the current GEM. Similarly, because the GEM does not account for optimal codon usage, the *in vivo* performance of non-native genes and proteins cannot be predicted.

The *E. coli* GEM has tilted the field of metabolic engineering towards advanced rational strain design by enabling researchers to explore a vast native and non-native genetic space in designing strains for improved metabolite production. More complete biochemical information will greatly aid metabolic engineering by allowing for genome-scale reconstructions that account for cellular functions beyond those accounted for in the metabolic models (e.g. regulation, expression, and enzyme kinetics). This implies the need for greater experimental method development to deduce the details of expression, post-

transcriptional and post-translational modifications, and enzyme kinetics. Advancements in pathway finding procedures to identify heterologous pathways that are not native to the host, and the techniques to optimize the expression of non-native enzymes, will expand the type of compounds that can be overproduced at an industrial scale. Overall, the ability of the *E. coli* GEM to aid systems level analysis for rational strain design will only continue to improve the speed with which viable production strains can be designed and constructed.

2. Biological discovery

There are aspects of bacterial functions that remain uncharacterized. Even in *E. coli*, the most studied and best-known bacterium, 34% of the genes have an unknown function [17]. In order to more efficiently expand our current understanding of cellular functions, an iterative workflow is needed that allows researchers to 1) account for what is known, 2) identify gaps in our knowledge, and 3) allow for the design of experiments to elucidate these gaps. The *E. coli* GEM has enabled the implementation of such a workflow to discover new features of microbial metabolism (Supplementary Table 4).

The function of uncharacterized open reading frames (ORFs) can be elucidated by comparing growth phenotypes from *in silico* model predictions of gene deletion mutants to *in vivo* experimental data (Figure 6). Discrepancies between GEM predictions and experimental results can point to where current knowledge is missing or where there are functional discrepancies. This, in turn, allows one to systematically formulate testable hypotheses. For example, incorrect predictions made for *talAB* mutants grown on xylose led the authors to

discover a novel pathway catalyzed by the gene products of *pfkA* and *fbaA* [35]. Various algorithms have been implemented to aid researchers in this process by parsing the vast number of biochemical pathways of metabolism to reconcile *in silico* growth predictions with experimental data [36-38]. These algorithms suggest network modifications (including assignment of enzymatic function to uncharacterized ORFs) that can then be confirmed by researchers *in vivo* [37]. For example, one study used a combination of graph-theory-based and comparative genomic analyses to identify *yneI* (*sad*) as the gene responsible for the NAD+/NADP+ -dependent succinate semialdehyde dehydrogenase, which the authors experimentally confirmed [39].

While the GEM-enabled workflow [36, 37, 40-42] has advanced our understanding of metabolism greatly, many aspects of bacterial metabolism are still waiting to be uncovered. One such aspect is the transcriptional regulation of metabolism. Researchers have attempted to integrate the transcriptional regulatory network (TRN) with the metabolic network to better understand and prediction regulation. For example, The TRN was used to elucidate changes in expression of oxygen regulators between oxic and anoxic conditions [40]. In another example, the TRN was used to confirm and refine the regulatory and functional assignment of various regulatory and metabolic genes, which included the novel finding that D-allose induces *rpiR* [36]. However, the Boolean formulation of the TRN regulatory rules only allows one to model regulatory interactions as either on or off. Consequently, complex regulatory interactions involving a multitude of transcription factors, binding constants, and environmental

dependencies along with post-transcriptional and post-translational modifications that may account for *in silico* and *in vivo* discrepancies cannot be identified using the current GEMs.

Efforts are underway to better understand bacterial regulation. Large-scale, genome-wide screens to deduce the function of transcriptional regulators and the development of new formalisms to account for and integrate transcriptional regulation in the model are in progress [43]. RNA sequencing-based technology will greatly assist researchers in elucidating the interactions of the transcriptional network by providing a richer data set than existing methods (e.g., RNA microarrays and ChIP-chip) [43]. Other post-transcriptional and post-translational modifications (these include, for instance, by small RNA at the transcript level or by phosphorylation, methylation, glycosylation, acetylation, or carboxylation, to name several, at the protein level) also contribute to the regulation of metabolic function in prokaryotes. Exemplary experimental efforts to better understand small RNA regulation of transcription [44], and the conservation of phosphorylation in serine, tyrosine, and threonine metabolism [45] have demonstrated that the link between gene expression and metabolite profiles are far more complex than once thought. Continued experimental efforts and improved computational efforts to deduce and model the complexities of regulation will be needed to expand the scope of biological discoveries that the current model can assist with.

3. Phenotypic functions

For simple organisms, the physiology of bacteria is remarkably complex. The diverse set of biochemical pathways in bacteria have conferred a vast phenotypic potential that have enabled them to thrive in a plethora of environments ranging from volcanic vents on the bottom of the ocean, clouds, glaciers, and the human gut. In order to understand this phenotypic potential, researchers have turned to GEMs to interpret and predict cellular phenotypes. Constraint-based modeling with GEMs [9] has allowed researchers to rapidly predict growth phenotypes in various genetic [46] and environmental [47] conditions, explore different objectives of microbial metabolism [48, 49] to examine the driving force behind cellular function, and better understand the suboptimal behavior of cells following perturbation [50, 51] and latent pathway activation [52].

When phenotypic predictions are made using the GEM, one finds a large solution space of potential phenotypes that would allow the organism to survive in a given genetic and environmental background. Many of these solutions of metabolic network usage that would allow for survival may not be observed under physiological conditions. Consequently, researchers have formulated ways to confine the solution space to more accurately represent the experimentally observed phenotype of the cell for a given growth condition by incorporating constraints. Regulatory control of the metabolic genes provides a means to constrain the allowable solution space by specifying what genes are active in the metabolic network for a given environmental condition. Researchers have been able to show that while over 50% of the flux distribution is constrained by metabolism in a given environmental condition, an additional 20% can be

attributed to the transcriptional regulatory network [53] (TRN). In addition, the TRN allows the organism to rapidly and efficiently adapt its metabolism to a wide range of environmental conditions by altering the expressed metabolic genes [54, 55]. Spatial constraints in the form of molecular crowding [56, 57], growth associated metabolite dilution [58], membrane occupancy [59], super coiling of the DNA [60], and indirect protein-protein interactions to facilitate the organization of enzymes in the cytoplasm [61] have shown promise for increasing the predictive power of the *E. coli* GEMs. For instance, a mechanistic constraint on the available space on the cytoplasmic membrane was introduced to better explain respiro-fermentation physiology [59]. Researchers have also shown that the physical structure of the DNA correlates more closely to the metabolic state than the regulatory network [60]. Thermodynamic analysis provides another means to constrain the solution space by removing infeasible reaction loops that violate energy balance [62, 63], by refining reaction directionality, and by defining allowable flux ranges via calculation of the *in vivo* change of Gibbs free energy of reactions [64-66]. Intimately tied to thermodynamic analysis is the integration of high-throughput metabolomics data. Several studies have incorporated metabolomics data into the *E. coli* GEM to better calculate the *in vivo* change in Gibbs free energy of reaction in order to better confine the feasible flux range of reactions and identify reactions that are potentially under allosteric regulation [67, 68].

The metabolic model can be used as a scaffold onto which high-throughput data types, such as fluxomic [69-71], transcriptomic [72, 73], and proteomic data [74], can be mapped to gain insight into context-specific phenotypes.

Fluxomics data can be used to directly compare intracellular flux distribution as predicted by constraint-based models [69, 70] using the GEM and can even be incorporated as additional constraints [71]. Recent work has also demonstrated that fluxomic data can be integrated into a computational framework to explain suboptimal behavior as a trade-off between near optimal growth under one condition, and the ability to quickly adapt to a new growth condition [75]. Transcriptomic data provides the experimentalist with a powerful means to decipher the phenotypic behavior of the cell due to its ability to qualitatively or quantitatively determine what genes are expressed by the cell under the given experimental conditions [73]. A combination of transcriptomic and proteomic data have also allowed researchers to better understand the physiology of adapted strains and the mechanism for this adaption [74].

While the *E. coli* GEM has aided our understanding of cellular metabolism, it has limitations. For instance, one should be aware that alternate optimal flux distributions of the cell may confound the researcher's ability to determine the true physiological state. This has been demonstrated when comparing fluxomic data to *in silico* predictions [70], and also in studies of adapted *E. coli* mutants [46, 76] where variations found in evolved replicates reflected the possible existence of multiple flux distributions that lead to equivalent growth phenotypes. The methods to predict cellular physiology present a user bias in the form of an objective function that must be validated for specific growth conditions [48]. The suboptimal state of the cell can also be predicted, but the most utilized method [50] provides little insight into the biological driving force for suboptimal performance

[77]. It appears instead that a Pareto optimal solution of multiple (and at times) conflicting objectives can better explain the biological significance of suboptimal behavior than any one method [75]. Incorporating thermodynamic constraints has been demonstrated to greatly reduce the solution space of metabolism. Unfortunately, the calculation of Gibbs free energy of reaction is hampered by the limited availability of experimentally determined standard Gibbs free energy of formation for a majority of the metabolites in the *E. coli* GEM. Therefore, the free energies of reaction can only be estimated [78].

Advancements in methods to obtain and integrate high quality omics data with the model will aid in overcoming many current limitations in accurately predicting phenotypic behavior. For example, genome-scale metabolomics is hampered by the biochemical diversity, range of physiological concentrations, and chemical liability of the species that comprise the intracellular metabolome. Consequently, multiple analytical platforms and well-tested analytical procedures [79] are needed to accurately assay the full metabolome, which is costly, time-consuming, and technically challenging. The enzyme kinetic and thermodynamic information that can be obtained from metabolomics can directly improve the accuracy of the metabolic model, and can also be correlated with gene expression profiles [80] to assist in unraveling the dynamics between transcriptome and metabolome. Another promising area is the formulation of a genome-scale isotope mapping model [81] for implementation with metabolic flux analysis. An expansion of metabolic flux analysis to the genome-scale and the ability to

determine the intracellular distribution of atoms beyond carbon will enhance our knowledge of *in vivo* flux states.

4. Biological network analysis

The metabolic reaction network is a highly complex, interwoven, and non-linear system that responds to environmental and genetic perturbations. In order to elucidate and understand the relationship between the network structure and function, many researchers have turned to network analysis. This exercise is mathematical in nature. In network analysis, biochemical reactions are transformed into a unipartite or bipartite graph, where the nodes and links take the form of metabolites and enzymatic reactions. Once formulated as a graph, the network can be sampled and explored using a variety of minimally biased mathematical and algorithmic methods to arrive at biologically insightful conclusions. The following paragraphs will focus on the most recent advances in biological network analysis; the reader is referenced to [18] and [31] for less recent examples not covered in the main text.

Much progress has been made in the analysis of link and node essentiality, whereby the consequences of removing a link (i.e., reaction) from the network is examined. Since links have varying degrees of dependence upon one another, one must look to higher order combinations of link removals to better understand the network properties of the *E. coli* GEMs. For example, synthetic lethals, which are defined as two genes whose independent deletion is not lethal, but simultaneous elimination is lethal, are often a consequence of network redundancy or parallel pathways [82]. The converse of synthetic lethals,

synthetic rescues, which are defined as a gene pair where the deletion of one of the genes is lethal, but the simultaneous deletion of both genes is non-lethal, can be used to rescue a nonviable single-gene deletion phenotype by rewiring the network in such a way as to compensate for the deleterious effect of the initial genetic perturbation [83, 84]. To illustrate, it has been shown that the over-expression of *udhA* improves the growth of *E. coli pgi* knockout strains on glucose minimal media [83]. Higher order epistatic interactions have also been analyzed to predict non-intuitive combinations of lethal and auxotrophic-inducing/rescuing gene-deletions[85].

It is important to emphasize that although the *E. coli* metabolic network is analogous to other interaction networks (e.g., the internet) and can be interrogated using network theory [86, 87], the *E. coli* network is unique in that it is a biological network that describes a highly evolved function. Network analysis can easily be taken out of context or provide little insight if the function of the metabolic network is not taken into account. For example, graph theory can deduce the topological properties of the model without providing any information about the underlying biology. A prerequisite of biologically meaningful network analysis is a biologically functional random network, to which one can compare the properties of the *E. coli* GEM [88, 89]. However, the *in vivo* experimental validation of such comparisons (i.e., to a random network) is infeasible. In addition, many of the network analysis methods become computationally challenged for large biochemical networks. Examples include elementary mode [90] and extreme pathway analyses [91], which have been for the most part limited to

small-scale networks due to the combinatorial explosion inherent to the methods [92, 93]. It should be noted that numerical efforts have been made to scale pathway analysis to genome-scale models by calculating only a subset of elementary modes [94]. Additionally, the *E. coli* metabolic network model is subject to iterative updates. Analyses obtained between older and newer models can lead to vastly different results. For instance, 81% of the coupling relations identified using flux coupling analysis changed between *i*JR904 and *i*AF1260 due to missing reactions in the older network model [95].

Network analysis using GEMs has largely been depicted as a strictly *in silico* undertaking. Recent progress, however, indicates that network analysis has practical applications. For example, the recent advances in elementary mode analysis can be readily extended to strain design [96] and non-native pathway finding procedures [97]. In another example, progress has been made in applying network analysis of the *E. coli* metabolic GEM to discover novel drug targets [98-100]. The *E. coli* GEM is particularly suitable for this application by enabling pharmaceutical researchers to analyze the complex interactions of the network as a whole to elucidate target links and nodes that would allow for complete system collapse or would severely cripple the network if removed. A potential viable workflow for antimicrobial drug discovery was recently presented [99]. The workflow invoked network analysis to identify novel antimicrobial targets combined with computational screening to identify inhibitory molecules against them followed by experimental validation [99]. This GEM-aided workflow could

reduce the expensive and time-consuming experimental methods in drug discovery [99], and is readily extendible to other bacterial species.

5. Bacterial evolution

The genomic content and phenotypic landscape of bacterial species are constantly adapting to meet the demands of the imposed environmental conditions. Adaption occurs via elimination of individual reactions by loss of function mutations, alterations in gene expression and enzyme capacity, alterations in enzyme kinetics, and through the addition of new reactions by horizontal gene transfer, gene duplication, and gain of function mutations. The *E. coli* GEM has proven most useful in modeling microbial evolution through elimination and addition of new metabolic network content, and acting as a scaffold to aid in the understanding of bacterial evolution.

Computational frameworks using GEMs have been employed to simulate bacterial evolution through random gene deletions. These studies have shown that there appears to be a conserved reaction set that is similar for organisms with similar lifestyles [101], which reflects the common enzymatic machinery required to metabolize specific carbon sources. It has also been shown that although genes are lost at random, the order in which genes are lost follows a coordinated and consistent pattern—40% of which can be accounted for by the metabolic model when compared to available phylogenic data [102]. The *E. coli* GEM also provides a context by which phylogeny data can be understood. Comparative genomics in the context of constraint-based modeling with the *E. coli* GEM has led researchers to assert that the dominant mechanism of bacterial

evolution in *E. coli* appears to be horizontal gene transfer. Horizontal gene transfer is highly dependent upon the genomic content of the organism [103, 104], and involves genes that are mostly environment-specific and located at the periphery of the metabolic network [105].

The *E. coli* GEM can act as a scaffold on which similar bacterial strains can be reconstructed, and their divergent evolutions understood. Due to the high standard of biochemical accuracy of the *E. coli* metabolic GEM (e.g., 97% of the included genetic content of the most recent *E. coli* GEM has been experimentally validated[17], many researchers have based the reconstruction of specific pathways or the entire organism on the reactions of the *E. coli* GEM (e.g., *Salmonella typhimurium* [106]). More recently, the pangenome of the species *E. coli* was reconstructed based on *i*AF1260, and used to generate 5 strain-specific GEMs that include commensal strain K-12 W3110, two enterohemorrhagic O157:H7 strains EDL933 and Sakai, and two uropathogenic strains CFT073 and UTI89 [107]. The study found that pathogenic *E. coli* appear to be more adapted to growth under anaerobic conditions than commensal *E. coli* [107]. The use of the *E. coli* GEM to rapidly construct strain-specific models will continue to increase, particularly as the cost of genome-sequencing of microbes continues to fall and the available number of sequenced and annotated strains continues to rise.

While the metabolic model allows a vast region of genotypic space to be explored in order to model and understand bacterial evolution, the space is currently limited to metabolic genes. Changes in the regulation of metabolism during evolution are not accounted for in GEMs. While horizontal gene transfer

and gene loss can be modeled up to the resolution of a core metabolic gene set [101], and the predicted gene loss order can be compared to evidence provided by comparative genomics [102], limitations remain in determining the precise genes and their exact loss, the location of mutations in the genome, and predicting their effect on the physiology of the organism. In addition, comparison of the evolutionary trajectories of different bacterial strains is hampered by the fact that strain-specific portions of the genomes remain largely uncharacterized.

The use of GEMs in modeling and understanding bacterial evolution will benefit from studies of adaptive laboratory evolution (ALE). ALE is an experimental procedure that introduces a selection pressure (e.g., fastest growth) in a controlled environmental setting that allows for a time-resolved depiction of changes in the organism's genome that occur during the process of adaption [108]. These changes can then be reintroduced and their effect on the organism's fitness studied [109-112]. The GEM provides a context for understanding these mutations by allowing the researcher to model the growth physiology of the adapted network.

6. Interspecies Interaction

There is a growing interest in better understanding host-pathogen interactions for the development of improved antimicrobials [113], the use of microbes for environmental remediation [114], and for understanding and manipulating the microbiome of the human gut for improved health [115]. Such applications would benefit greatly from a platform that would allow for the prediction and simulation of biological interactions. The *E. coli* GEM provides

such a platform and has been successful in modeling the exchange of metabolites (Figure 7) between different cell types (e.g., microbial species) and environmental conditions [116-118].

Although interaction classes of interspecies interactions have been established, the mathematical formalism to model these interaction classes using GEMs has arisen only recently [116-120]. Researchers have modeled host/pathogen interactions by directly incorporating pathogenic reactions into the stoichiometry of the host reaction network (e.g., to account for viral amino acid and nucleotide synthesis [116]). Concatenated and joint stoichiometric models have been employed to study the exchange of metabolites between species (e.g. co-cultures [118, 119]) and the environment [117]. In contrast to multi-cellular stoichiometric models that assume that the collection of microbes seeks to maximize the collective biomass, researchers have developed multi-competitor metabolic models to describe microbial communities where each member seeks to maximize their own biomass [120]. Together, these studies have found that the combined metabolisms of multiple species can utilize the capabilities of the environment better than a single species. Furthermore, while not every metabolic interaction is beneficial, metabolic interactions necessitate community formation.

Much work remains in the field of modeling interspecies interactions. For example, the measurement of metabolites exchanged between cells, needed to validate the accuracy of the *in silico* predictions, presents a strong technical challenge. In addition, the effect of biological interactions on regulatory elements

is not yet accounted for. For instance, the extent to which other strains and the environment influence the regulation of metabolism (e.g., through quorum sensing) is unclear. Also, most of the large community models do not differentiate the genetic content between individual species nor account for their spatial organization. Consideration of regulation, individual genetic content, and spatial organization will be needed to more accurately model and predict community level biological processes [121].

The advent of single-cell sequencing technology [122] and other single-cell assays [123] will benefit the study of interspecies interactions with GEMs. Such a 'bottom-up' approach would allow for the characterization of individual biological entities through, for example, genomics and transcriptomics of individual species. From a 'top-down' approach, the total interaction of microbial communities can be characterized through genome-scale-omic data. Genome sequencing and reconstruction of other bacterial species through manual curation or automatic reconstruction (e.g., model SEED [3]), and the mapping of metabolite and reaction identifiers between reconstructions (e.g., RxnMet [124]) will expand the number of species interactions that can be simulated together. Based on genome-scale-omic data that correspond with single cell data, developments in bioinformatics approaches that establish the relationships between individual biological entities, and the growing number of reconstructions will allow researchers to piece together the contribution of each member on the biological community in order to generate complete community level models.

In closing: What is likely for the future of GEMs

The number of applications focused on *E. coli* that have utilized the metabolic GEM have grown in size and scope [18]. This review distilled into six categories the approximately 200 studies that have appeared in peer-reviewed manuscripts over the past 12 years. In each category, key examples and success stories were summarized and presented. In addition, we critically analyzed the current status of applications using the *E. coli* metabolic GEM to demonstrate what the model can and cannot do, and discussed the developments needed to overcome current limitations; as summarized in Table 1. To help summarize the impact of GEM-aided analyses thus far, Supplementary Figure 2 highlights studies that have made a significant contribution to the *E. coli* GEM in particular, and our understanding of microbial metabolism, in general. Researchers are now able to complete the systems biology workflow and generate new biological knowledge with the help of the *E. coli* GEM (Fig 4). It should be emphasized that while this review has focused on the metabolic network, it is but one of several networks actively at work inside the cell [2].

The GEM of *E. coli* will continue to expand as more cellular processes are mechanistically detailed and added to the organized GEM structure. The next significant increase in the applications of an *E. coli* GEM will likely come from mechanistically incorporating and integrating protein synthesis with metabolism. The integration of the transcriptional and translational machinery on the genome scale [125, 126]has now been completed. The operon structure that accounts for

cellular regulation will follow protein synthesis as the next logical step of GEM expansion. The incorporation of DNA structure and transcription binding as a ready-to-compute biochemical network in a mathematical format would overcome the limitations presented by the current Boolean formulation of the TRN, and allow for complex regulatory interactions to be mechanistically modeled and predicted. It is conceivable that DNA synthesis, post-translational modifications, and other cellular processes that involve biochemical interactions that can be described by a biochemical interaction network can also be incorporated into GEMs. In brief, what lies ahead for GEMs is the iterative expansion to include other cellular processes beyond metabolism with the aid of omics data and the mathematical formalisms to model them (Fig 5). It is unclear which high-throughput data types and algorithms will be the major drivers for many of the applications enabled by GEMs with an expanded scope. However, it is clear that modeling with such expansive networks, whose components will carry activities across many orders of magnitude, will require greater computational accuracy and power given their size. Furthermore, the payoff for this increased complexity will be more accurate phenotype predictions after initial validation and gap filling is performed. GEM expansion will be a substantial but worthwhile endeavor that will unite many diverse aspects of microbiology and move the community closer to the ultimate goal of establishing a comprehensive mechanistic understanding of the genotype-phenotype relationship of microbes.

Acknowledgments

Chapter 2, in full, is a reformatted reprint of “Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli.” McCloskey D, Palsson BO, Feist AM. Mol Syst Biol. 2013;9:661. doi: 10.1038/msb.2013.18. The dissertation/thesis author was the primary investigator and author of the paper.

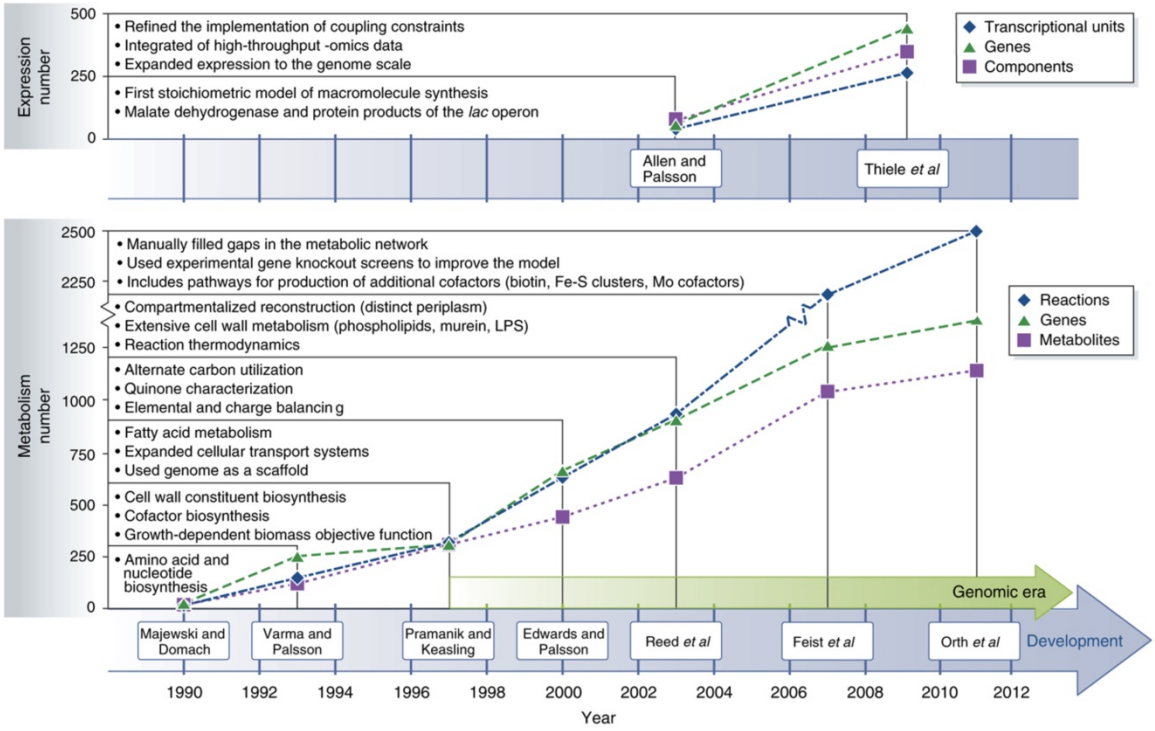

Figure 2.1: History of the *E. coli* expression and metabolic reconstructions. Shown in the upper portion of the graph are 2 milestone efforts contributing to the reconstruction of the *E. coli* transcription and translation network, and shown in the bottom portion of the graph are seven milestone efforts contributing to the reconstruction of the *E. coli* metabolic network. For each of the two reconstructions shown [126, 127] in the upper graph, the number of included transcription units (blue diamonds), genes (green triangles) and components (purple squares) are displayed. For each of the seven reconstructions shown [10, 11, 17, 128-131] in the bottom graph, the number of included reactions (blue diamonds), genes (green triangles) and metabolites (purple squares) are displayed. Also listed is noteworthy content expansion that each successive reconstruction provided over previous efforts. For example, Varma & Palsson [11, 12] included amino acid and nucleotide biosynthesis pathways in addition to the content that Majewski & Domach [10] characterized. The start of the genomic era [15] marked a significant increase in included components for successive iterations of the network reconstruction. The significant increase in the number of reactions in 2007 [130] was in large part due to the removal of many lumped reactions, which were often included for lipid and cell wall biosynthesis in earlier metabolic reconstructions. Thiele *et. al.* [126] expanded the initial work of Allen & Palsson [127] by increasing the scope of the transcription and translation network from a few example pathways to all known genes involved in protein synthesis (i.e., expression). Not included on the timeline is a metabolic reconstruction based upon Reed *et al.*, that was modified to include additional reactions from the KEGG [132] database and incorporated into the MetaFluxNet software package[133].

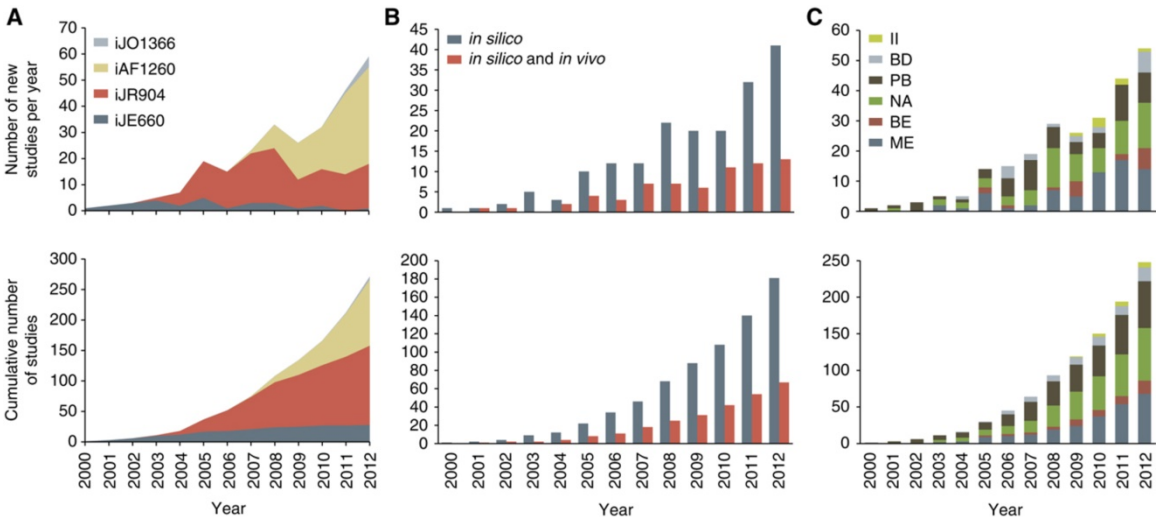

Figure 2.2: The detailed usage of the *E. coli* metabolic GEM over time. The cumulative and new number of studies published per year separated according to a) the metabolic reconstruction used [17, 128-130], b) *in silico* (i.e., strictly computational prediction) or combined *in silico & vivo* (i.e., computational usage of the model and experimental validation or data generation guided by the model), and c) the application category of the study. ME: metabolic engineering, BE: studies of evolutionary processes, NA: analysis of network properties, PB: prediction of cellular phenotypes, BD: model-driven discovery, II: interspecies interaction.

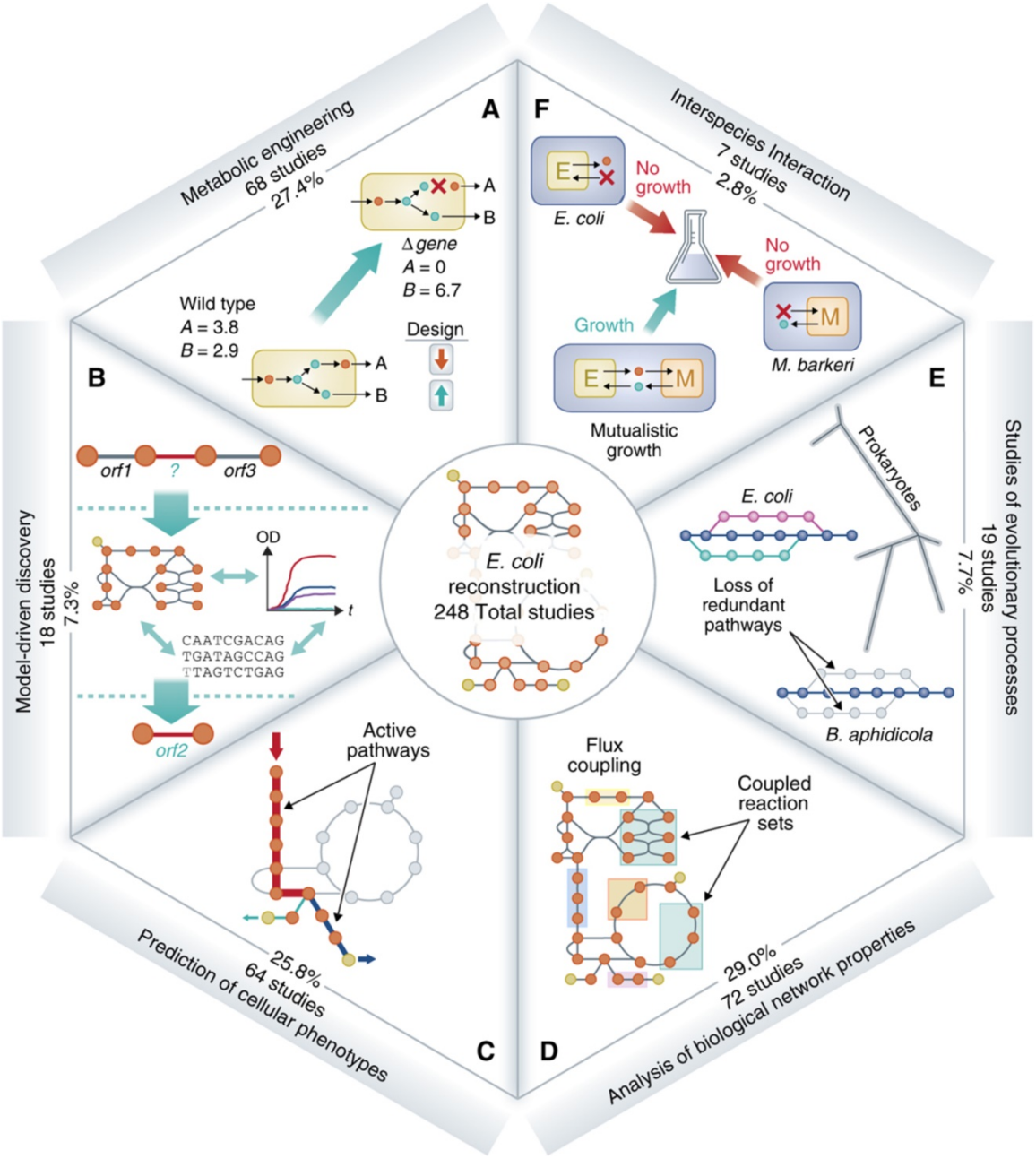

Figure 2.3: Six categories of uses and number of studies for each use of the *E. coli* metabolic GEM. The original five categories defined in 2008 [18] include a) metabolic engineering, b) model-driven discovery, c) prediction of cellular phenotypes, d) Analysis of biological network properties, and e) studies of evolutionary processes. A new category has been added, f) interspecies interaction. The addition of this category signifies a growing trend in the field to explore the interaction of the *E. coli* metabolic network with other organisms and across different environmental conditions. Specifically, studies have explored host/pathogen interactions [116], co-cultures [118-120], ecology [117], and chemotaxis [134]. The number of studies in this category is expected to increase as the interest in understanding the complexities of microbial interactions and ecosystems continues to grow. The complete lists of the studies for each category are included in Supplementary Table 1.

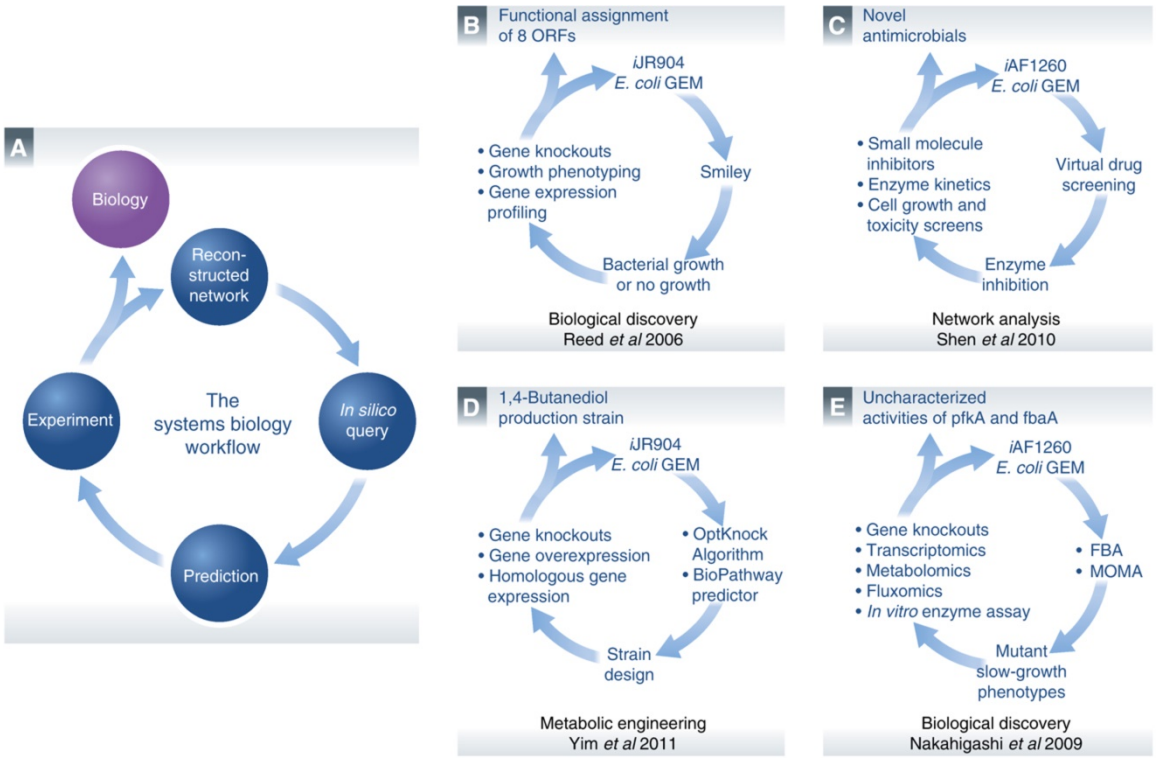

Figure 2.4: Iterative workflows. A) A generic network reconstruction and model driven systems biology workflow is a cyclic path that iterates between *in silico* predictions and *in vivo* observations. This general process has been followed in some of the more influential studies presented in this review. DNA sequencing and bibliomic data can be used to reconstruct and translate a biological system into a mathematical structure. Other omics data types that have been generated can be interpreted in the context of a reconstruction and computational model to analyze organism functions under specific conditions. This information becomes a *de facto* knowledgebase that can be queried through a consortium of analytical methods. The aim of these methods is to hypothesize answers to complex biological questions that can often be non-intuitive or not readily apparent. Experiments can then be designed to test these predictions in order to either confirm GEM-derived explanations or move researchers one iteration closer to the answer. Studies that have successfully iterated through the *E. coli* GEM workflow that are presented as examples include B) Reed et al., 2006 [37], C) Shen et al., 2010 [99], D) Yim et al., 2011 [25], and E) Nakahigashi et al., 2009 [35].

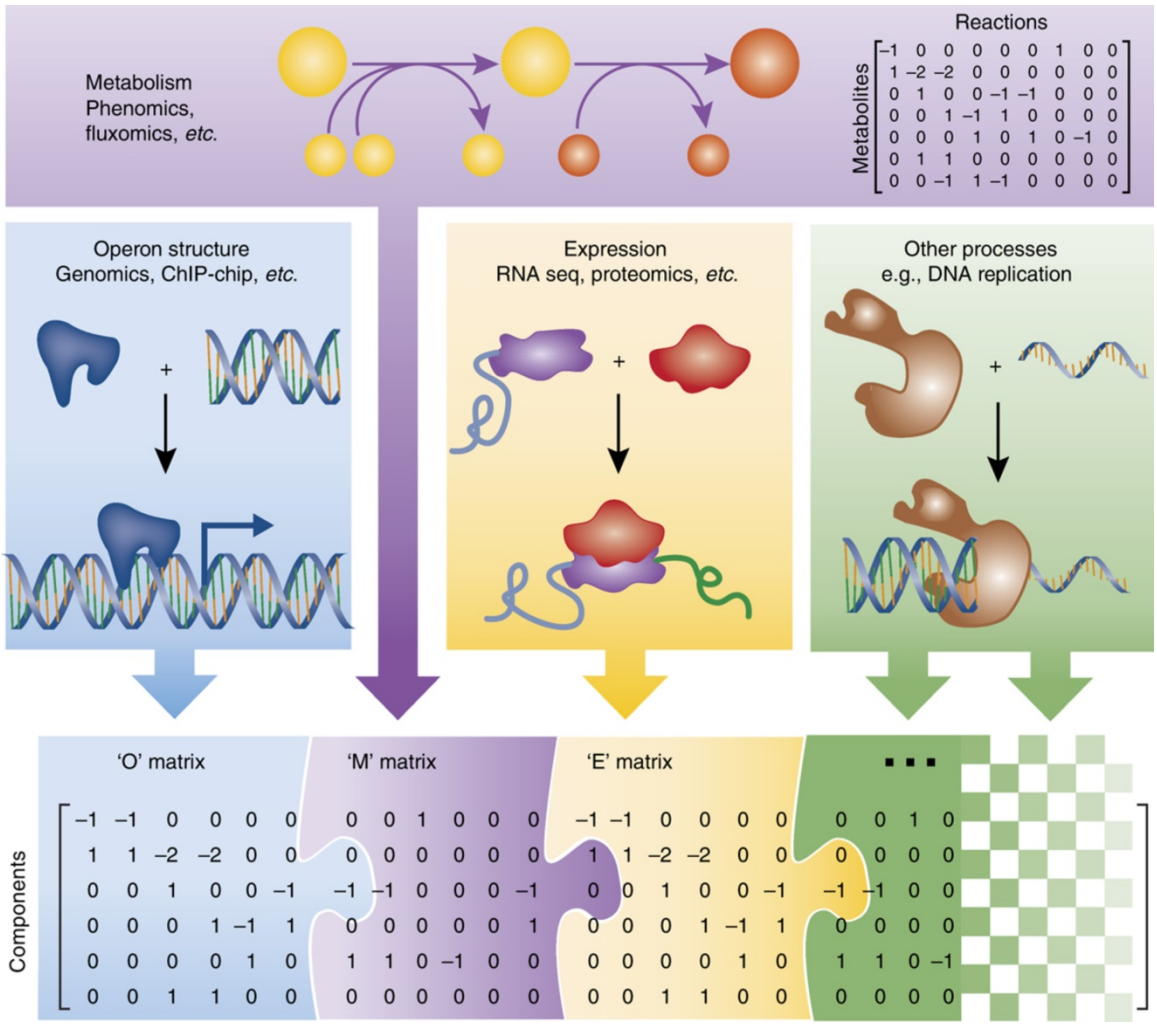

Figure 2.5: The future of the *E. coli* GEM. The most widely-used *E. coli* reconstruction accounts only for metabolism (the "M" matrix) [2]. However, efforts are currently underway to integrate the operon structure that determines cellular regulation (the "O" matrix), the transcriptional and translational machinery allowing for the expression of proteins (the "E" matrix) [126], and other cellular processes (e.g., DNA replication, post-translational modifications, etc.) with metabolism. The integration of these cellular processes, supported by high-throughput data types, into a single mathematical model, will allow researchers to more accurately compute complex phenotypes, and will guide the discovery of unknown aspects of cellular functions beyond that of just metabolism.

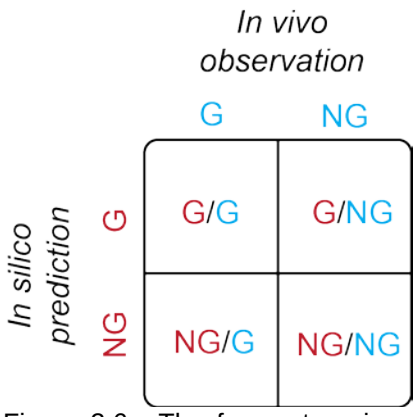

Figure 2.6: The four categories of growth phenotypes. Growth phenotypes from single-gene deletion mutants based on *in silico* model predictions can be compared to *in vivo* experimental data to elucidate or confirm function of ORFs. The results of growth phenotyping studies can be classified into four categories: 1) Growth/Growth (G/G): the model and experimental data show growth, 2) Growth/No Growth (G/NG): the model predicts growth, but the experimental data indicates no growth, 3) No Growth/Growth (NG/G): the model predicts no growth, but the experimental data indicates growth, and 4) Growth/No Growth (NG/NG): the model and experimental data show no growth. The G/NG case indicates that the model over-estimates the metabolic capabilities of the organism, while the NG/G case indicates that the model under-estimates the metabolic capabilities of the organism. Metabolic over-predictions are commonly caused by reactions that are absent *in vivo*, reactions that are down-regulated or inhibited under a specific environmental condition, or the biomass formulation is includes an erroneous metabolite. Metabolic under-predictions often represent knowledge gaps in that the model does not account for an unknown isozyme, parallel pathway, or some other functionality of the organism. G/G can be regarded as a consistency check and NG/NG can be regarded as a form of model validation.

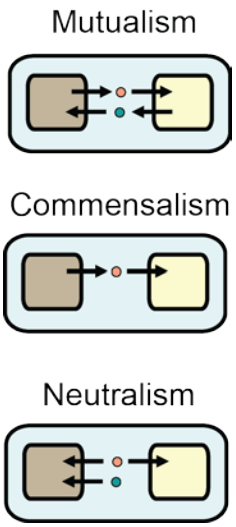

Figure 2.7: Microbial interactions as defined by metabolite exchange. A microbial interaction can be defined and modeled as an exchange of molecules between species in a given environment. There are three main types of interaction classes between microbial species: 1. Mutualism, also known as syntrophy or symbiosis, is where each organism produces an essential metabolite needed to support growth by the other organism. 2. Commensalism is where only one organism depends on the other for the production of an essential nutrient to support growth. A special case of commensalism, known as parasitism, is when the organism providing the essential nutrient comes at the cost of reduced fitness. Host/pathogen interactions are a type of parasitism. 3. Neutralism is where each organism can sustain growth in a given environment without the presence of the other organism. Since each species are consuming the same resources, competition can arise between the organisms.

Table 2.1: Strengths and limitations of the metabolic GEM applications.

Application	What the model can do: *Strengths of the E. coli GEM*	What the model cannot do: *Areas for future progress*
Metabolic Engineering	• Gene deletion (combinatorial) • Gene addition • Gene over-and under-expression • Rapidly test the systemic effects of heterologous pathway additions • Design biomarkers/biosensors for characteristic function • Determine media supplementation strategies • Map high-throughput data to identify bottlenecks • Design strains through evolution	• Limited coverage of molecular biology • Predicting the effects of perturbations to regulatory elements • Predicting allosteric inhibition • There is no explicit representation of metabolite concentrations • Account for enzyme kinetics • Cannot accurately predict the performance of non-native genes/proteins in *E. coli*
Biological Discovery	• Predict growth on different carbon sources / media conditions • Guide the functional assignment of network gaps • Guide the discovery of previously uncharacterized gene product functions (graph theory analysis) • Guide the re-annotations of incorrectly annotated genes • Connect orphan metabolites to known reactions	• Predict the regulation of isozymes/parallel pathways • Predict enzyme promiscuity • Predictive power is inherently limited because the model is not complete in scope • Predict the expression of genes • Predict the functional state of proteins (e.g. post-translational modification)
Phenotypic Behavior	• Predict optimal cellular behavior • Understand energetics and occurrence of suboptimal behavior • Infer impact of regulation • Provide a context for which experimental data can be interpreted • Predict and understand absolute and conditional gene essentiality • Predict and understand shifts in growth conditions	• Differentiate between computed alternate optimal flux distributions of the cell *a priori* • Explain the reasons for suboptimal performance *a priori* • Provide a framework for incorporating additional regulatory interactions that are currently under development
Network Analysis	• Evaluate metabolic networks from a systems view through node and link dependencies, essentialities, overall network robustness • Describe the complex interactions of the components of the metabolic network • Evaluate modularity of function • Evaluate regulation based on network structure	• Does not always include the biological mechanisms behind the network connections • Few predictions can be experimentally validated
Bacterial Evolution	• Predict essential genes • Predict the end-point of evolution • Understand the basis for epistatic interactions and mutational effects • Provide insights into evolutionary trajectories	• Account for changes in regulatory elements • Predict the time-course of evolution • Predict location of mutations in the genome • Predict the effects of mutations in the genome • Account for strain-specific genomic differences
Interspecies Interaction	• Model the exchange of metabolites • Analyze high-throughput data from different strains • Determine the cost/benefit ratio for different types of commensalism	• Model interactions that affect metabolic regulation • Inability to measure flux exchange in multi cell-type systems • There are still too many unknowns to accurately build an interactions network • Limited ability to define individual genetic content in large communities • Limited spatial knowledge in large communities

References:

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

2. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. & Palsson, B.O. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* **7**, 129-143 (2009).

3. Henry, C.S. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* **28**, 977-982 (2010).

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

5. Palsson, B. Metabolic systems biology. *FEBS Lett* **583**, 3900-3904 (2009).

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

7. Pfau, T., Christian, N. & Ebenhoh, O. Systems approaches to modelling pathways and networks. *Brief Funct Genomics* **10**, 266-279 (2011).

8. Orth, J.D., Thiele, I. & Palsson, B.O. What is flux balance analysis? *Nat Biotechnol* **28**, 245-248 (2010).

9. Palsson, B. Systems biology : properties of reconstructed networks. (Cambridge University Press, New York; 2006).

10. Majewski, R.A. & Domach, M.M. Simple constrained-optimization view of acetate overflow in E. coli. *Biotechnol Bioeng* **35**, 732-738 (1990).

11. Varma, A., Boesch, B.W. & Palsson, B.O. Biochemical production capabilities of Escherichia coli. *Biotechnol Bioeng* **42**, 59-73 (1993).

12. Varma, A., Palsson, Bernhard O. Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns. *J. theor. Biol.* **165**, 503-522 (1993).

13. Varma, A. & Palsson, B.O. Metabolic Flux Balancing: Basic concepts, Scientific and Practical Use. *Nat Biotechnol* **12**, 994-998 (1994).

14. Varma, A. & Palsson, B.O. Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic by-Product Secretion in Wild-Type Escherichia-Coli W3110. *Appl Environ Microb* **60**, 3724-3731 (1994).

15. Blattner, F.R.. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-1462 (1997).

16. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640-6645 (2000).

17. Orth, J.D.. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* **7**, 535 (2011).

18. Feist, A.M. & Palsson, B.O. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* **26**, 659-667 (2008).

19. Sweetlove, L.J. & Ratcliffe, R.G. Flux-balance modeling of plant metabolism. *Front Plant Sci* **2**, 38 (2011).

20. Osterlund, T., Nookaew, I. & Nielsen, J. Fifteen years of large scale metabolic modeling of yeast: developments and impacts. *Biotechnol Adv* **30**, 979-988 (2012).

21. Lee, J.W.. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat Chem Biol* **8**, 536-546 (2012).

22. Fong, S.S.. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

23. Fowler, Z.L., Gikandi, W.W. & Koffas, M.A.G. Increased Malonyl Coenzyme A Biosynthesis by Tuning the Escherichia coli Metabolic Network and Its Application to Flavanone Production. *Appl. Environ. Microbiol.* **75**, 5831-5839 (2009).

24. Lee, K.H., Park, J.H., Kim, T.Y., Kim, H.U. & Lee, S.Y. Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol* **3**, 149 (2007).

25. Yim, H.. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* **7**, 445-452 (2011).

26. Xu, P., Ranganathan, S., Fowler, Z.L., Maranas, C.D. & Koffas, M.A. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab Eng* **13**, 578-587 (2011).

27. Jung, Y.K., Kim, T.Y., Park, S.J. & Lee, S.Y. Metabolic engineering of Escherichia coli for the production of polylactic acid and its copolymers. *Biotechnol Bioeng* **105**, 161-171 (2010).

28. Patil, K.R., Rocha, I., Forster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6**, 308 (2005).

29. Kim, J. & Reed, J.L. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol* **4**, 53 (2010).

30. Burgard, A.P., Pharkya, P. & Maranas, C.D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering* **84**, 647-657 (2003).

31. Oberhardt, M.A., Palsson, B.O. & Papin, J.A. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5**, 320 (2009).

32. Copeland, W.B.. Computational tools for metabolic engineering. *Metab Eng* **14**, 270-280 (2012).

33. Chemler, J.A., Fowler, Z.L., McHugh, K.P. & Koffas, M.A.G. Improving NADPH availability for natural product biosynthesis in Escherichia coli by metabolic engineering. *Metabolic Engineering* **12**, 96-104 (2010).

34. Alper, H., Miyaoku, K. & Stephanopoulos, G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* **23**, 612-616 (2005).

35. Nakahigashi, K.. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol* **5**, 306 (2009).

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

37. Reed, J.L.. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* **103**, 17480-17484 (2006).

38. Satish Kumar, V., Dasika, M. & Maranas, C. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).

39. Fuhrer, T., Chen, L., Sauer, U. & Vitkup, D. Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in Escherichia coli. *J Bacteriol* **189**, 8073-8078 (2007).

40. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92-96 (2004).

41. Joyce, A.R.. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. *J Bacteriol* **188**, 8259-8271 (2006).

42. Holm, A.K.. Metabolic and Transcriptional Response to Cofactor Perturbations in Escherichia coli. *Journal of Biological Chemistry* **285**, 17498-17506 (2010).

43. Cho, B.-K., Palsson, B. & Zengler, K. Deciphering the regulatory codes in bacterial genomes. *Biotechnology Journal* **6**, 1052-1063 (2011).

44. Wassarman, K.M. & Saecker, R.M. Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science* **314**, 1601-1603 (2006).

45. Macek, B.. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics* **7**, 299-307 (2008).

46. Fong, S.S. & Palsson, B.O. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet* **36**, 1056-1058 (2004).

47. Ibarra, R.U., Edwards, J.S. & Palsson, B.O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186-189 (2002).

48. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol* **3**, 119 (2007).

49. Ow, D.S.-W., Lee, D.-Y., Yap, M.G.-S. & Oh, S.K.-W. Identification of cellular objective for elucidating the physiological state of plasmid-bearing Escherichia coli using genome-scale in silico analysis. *Biotechnology Progress* **25**, 61-67 (2009).

50. Segre, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99**, 15112-15117 (2002).

51. Link, H., Anselment, B. & Weuster-Botz, D. Rapid media transition: An experimental approach for steady state analysis of metabolic pathways. *Biotechnology Progress* **26**, 1-10 (2010).

52. Nishikawa, T., Gulbahce, N. & Motter, A.E. Spontaneous reaction silencing in metabolic optimization. *PLoS Comput Biol* **4**, e1000236 (2008).

53. Shlomi, T., Eisenberg, Y., Sharan, R. & Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology* **3**, 101 (2007).

54. Samal, A. & Jain, S. The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC systems biology* **2**, 21 (2008).

55. Barrett, C.L., Herring, C.D., Reed, J.L. & Palsson, B.O. The global transcriptional regulatory network for metabolism in *Escherichia coli* attains few dominant functional states. *Proc Natl Acad Sci U S A* **102**, 19103-19108 (2005).

56. Beg, Q.K.. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proc Natl Acad Sci U S A* **104**, 12663-12668 (2007).

57. Vazquez, A.. Impact of the solvent capacity constraint on *E. coli* metabolism. *BMC systems biology* **2**, 7 (2008).

58. Benyamini, T., Folger, O., Ruppin, E. & Shlomi, T. Flux balance analysis accounting for metabolite dilution. *Genome Biology* **11**, R43 (2010).

59. Zhuang, K., Vemuri, G.N. & Mahadevan, R. Economics of membrane occupancy and respiro-fermentation. *Mol Syst Biol* **7**, 500 (2011).

60. Sonnenschein, N., Geertz, M., Muskhelishvili, G. & Hutt, M.-T. Analog regulation of metabolic demand. *BMC systems biology* **5**, 40 (2011).

61. Perez-Bercoff, A., McLysaght, A. & Conant, G.C. Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Mol Biosyst* **7**, 3056-3064 (2011).

62. Ederer, M. & Gilles, E.D. Thermodynamically feasible kinetic models of reaction networks. *Biophysical journal* **92**, 1846-1857 (2007).

63. Fleming, R.M.T., Thiele, I. & Nasheuer, H.P. Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to Escherichia coli. *Biophysical Chemistry* **145**, 47-56 (2009).

64. Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* **92**, 1792-1805 (2007).

65. Kümmel, A., Panke, S. & Heinemann, M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* **7** (2006).

66. Flamholz, A., Noor, E., Bar-Even, A. & Milo, R. eQuilibrator--the biochemical thermodynamics calculator. *Nucleic Acids Res* **40**, D770-775 (2012).

67. Zamboni, N., Kummel, A. & Heinemann, M. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9**, 199 (2008).

68. Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. & Shlomi, T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **26**, i255-i260 (2010).

69. Herrgard, M.J., Fong, S.S. & Palsson, B.O. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* **2**, e72 (2006).

70. Chen, X., Alonso, A.P., Allen, D.K., Reed, J.L. & Shachar-Hill, Y. Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. *Metabolic Engineering* **13**, 38-48 (2011).

71. Choi, H.S., Kim, T.Y., Lee, D.Y. & Lee, S.Y. Incorporating metabolic flux ratios into constraint-based flux analysis by using artificial metabolites and converging ratio determinants. *Journal of biotechnology* **129**, 696-705 (2007).

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

73. Portnoy, V.A.. Deletion of Genes Encoding Cytochrome Oxidases and Quinol Monooxygenase Blocks the Aerobic-Anaerobic Shift in Escherichia coli K-12 MG1655. *Appl. Environ. Microbiol.* **76**, 6529-6540 (2010).

74. Lewis, N.E.. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology* **6**, 390 (2010).

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

76. Charusanti, P.. Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. *PLoS Genet* **6**, e1001186 (2010).

77. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* **102**, 7695-7700 (2005).

78. Jankowski, M.D., Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J* **95**, 1487-1499 (2008).

79. Buscher, J.M., Czernik, D., Ewald, J.C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal Chem* **81**, 2135-2143 (2009).

80. Jozefczuk, S.. Metabolomic and transcriptomic stress response of Escherichia coli. *Mol Syst Biol* **6**, 364 (2010).

81. Ravikirthi, P., Suthers, P.F. & Maranas, C.D. Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and Bioengineering* **108**, 1372-1382 (2011).

82. Ghim, C.M., Goh, K.I. & Kahng, B. Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. *Journal of Theoretical Biology* **237**, 401-411 (2005).

83. Kim, D.-H. & Motter, A.E. Slave nodes and the controllability of metabolic networks. *New Journal of Physics* **11**, 113047 (2009).

84. Motter, A.E., Gulbahce, N., Almaas, E. & Barabasi, A.L. Predicting synthetic rescues in metabolic networks. *Molecular systems biology* **4**, 168 (2008).

85. Suthers, P.F., Zomorrodi, A. & Maranas, C.D. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol* **5**, 301 (2009).

86. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. & Barabasi, A.L. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839-843 (2004).

87. Samal, A., Wagner, A. & Martin, O.C. Environmental versatility promotes modularity in genome-scale metabolic networks. *BMC Syst Biol* **5**, 135 (2011).

88. Samal, A. & Martin, O.C. Randomizing Genome-Scale Metabolic Networks. *Plos One* **6**, e22295 (2011).

89. Basler, G., Grimbs, S., Ebenhoh, O., Selbig, J. & Nikoloski, Z. Evolutionary significance of metabolic network properties. *J R Soc Interface* **9**, 1168-1176 (2012).

90. Schuster, S., Dandekar, T. & Fell, D.A. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology* **17**, 53-60 (1999).

91. Schilling, C.H., Letscher, D. & Palsson, B.O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology* **203**, 229-248 (2000).

92. Klamt, S. & Stelling, J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* **29**, 233-236 (2002).

93. Yeung, M., Thiele, I. & Palsson, B.O. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**, 363 (2007).

94. Wessely, F.. Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Mol Syst Biol* **7**, 515 (2011).

95. Marashi, S.-A. & Bockmayr, A. Flux coupling analysis of metabolic networks is sensitive to missing reactions. *Biosystems* **103**, 57-66 (2011).

96. de Figueiredo, L.F.. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **25**, 3158-3165 (2009).

97. Larhlimi, A., Basler, G., Grimbs, S., Selbig, J. & Nikoloski, Z. Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics* **28**, i502-i508 (2012).

98. Kim, T.Y., Kim, H.U. & Lee, S.Y. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metabolic Engineering* **12**, 105-111 (2010).

99. Shen, Y.. Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc Natl Acad Sci U S A* **107**, 1082-1087 (2010).

100. Plaimas, K.. Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC systems biology* **2**, 67 (2008).

101. Pal, C.. Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667-670 (2006).

102. Yizhak, K., Tuller, T., Papp, B. & Ruppin, E. Metabolic modeling of endosymbiont genome reduction on a temporal scale. *Mol Syst Biol* **7**, 479 (2011).

103. Pal, C., Papp, B. & Lercher, M.J. Horizontal gene transfer depends on gene content of the host. *Bioinformatics* **21 Suppl 2**, ii222-ii223 (2005).

104. Notebaart, R., Kensche, P., Huynen, M. & Dutilh, B. Asymmetric relationships between proteins shape genome evolution. *Genome Biology* **10**, R19 (2009).

105. Pal, C., Papp, B. & Lercher, M.J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**, 1372-1375 (2005).

106. AbuOun, M.. Genome Scale Reconstruction of a Salmonella Metabolic Model. *Journal of Biological Chemistry* **284**, 29480-29488 (2009).

107. Baumler, D., Peplinski, R., Reed, J., Glasner, J. & Perna, N. The evolution of metabolic networks of E. coli. *BMC systems biology* **5**, 182 (2011).

108. Conrad, T.M., Lewis, N.E. & Palsson, B.O. Microbial laboratory evolution in the era of genome-scale science. *Molecular systems biology* **7**, 509 (2011).

109. Herring, C.D.. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* **38**, 1406-1412 (2006).

110. Applebee, M.K., Joyce, A.R., Conrad, T.M., Pettigrew, D.W. & Palsson, B.O. Functional and metabolic effects of adaptive glycerol kinase (GLPK) mutants in Escherichia coli. *J Biol Chem* **286**, 23150-23159 (2011).

111. Lee, D.H. & Palsson, B.O. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microb* **76**, 4158-4168 (2010).

112. Conrad, T.M.. Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol* **10**, R118 (2009).

113. Lebeis, S.L. & Kalman, D. Aligning Antimicrobial Drug Discovery with Complex and Redundant Host-Pathogen Interactions. *Cell Host & Microbe* **5**, 114-122 (2009).

114. Singh, J.S., Abhilash, P.C., Singh, H.B., Singh, R.P. & Singh, D.P. Genetically engineered bacteria: An emerging tool for environmental remediation and future research perspectives. *Gene* **480**, 1-9 (2011).

115. Walter, J., Britton, R.A. & Roos, S. Host-microbial symbiosis in the vertebrate gastrointestinal tract and the Lactobacillus reuteri paradigm. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4645-4652 (2011).

116. Jain, R. & Srivastava, R. Metabolic investigation of host/pathogen interaction using MS2-infected Escherichia coli. *BMC systems biology* **3**, 121 (2009).

117. Klitgord, N. & Segrè, D. Environments that Induce Synthetic Microbial Ecosystems. *PLoS Comput Biol* **6**, e1001002 (2010).

118. Wintermute, E.H. & Silver, P.A. Emergent cooperation in microbial metabolism. *Mol Syst Biol* **6**, 407 (2010).

119. Hanly, T.J. & Henson, M.A. Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnology and Bioengineering* **108**, 376-385 (2011).

120. Tzamali, E., Poirazi, P., Tollis, I.G. & Reczko, M. A computational exploration of bacterial metabolic diversity identifying metabolic

interactions and growth-efficient strain communities. *BMC systems biology* **5**, 167 (2011).

121. Zengler, K. & Palsson, B.O. A road map for the development of community systems (CoSy) biology. *Nat Rev Microbiol* **10**, 366-372 (2012).

122. Tang, F.. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).

123. Taniguchi, Y.. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533-538 (2010).

124. Kumar, A., Suthers, P.F. & Maranas, C.D. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics* **13**, 6 (2012).

125. Thiele, I.. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. *PLoS ONE* **7**, e45635 (2012).

126. Thiele, I., Jamshidi, N., Fleming, R.M. & Palsson, B.O. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* **5**, e1000312 (2009).

127. Allen, T.E. & Palsson, B.O. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J Theor Biol* **220**, 1-18 (2003).

128. Edwards, J.S. & Palsson, B.O. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences* **97**, 5528-5533 (2000).

129. Reed, J., Vo, T., Schilling, C. & Palsson, B. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology* **4**, R54 (2003).

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

131. Pramanik, J. & Keasling, J.D. Stoichiometric model of Escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition

and mechanistic energy requirements. *Biotechnology and Bioengineering* **56**, 398-421 (1997).

132. Kanehisa, M.. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-484 (2008).

133. Lee, S.Y.. Systems-level analysis of genome-scale in silico metabolic models using MetaFluxNet. *Biotechnol Bioproc E* **10**, 425-431 (2005).

134. Kugler, H., Larjo, A. & Harel, D. Biocharts: a visual formalism for complex biological systems. *Journal of The Royal Society Interface* **7**, 1015-1024 (2010).

CHAPTER 3:

A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in *E. coli* K-12 MG1655 that is biochemically and thermodynamically consistent

Abstract

The advent of model-enabled workflows in systems biology allows for the integration of experimental data types with genome-scale models to discover new features of biology. This work demonstrates such a workflow, aimed at establishing a metabolomics platform applied to study the differences in metabolomes between anaerobic and aerobic growth of *Escherichia coli*. Constraint-based modeling was utilized to deduce a target list of compounds for downstream method development. An analytical and experimental methodology was developed and tailored to the compound chemistry and growth conditions of interest. This included the construction of a rapid sampling apparatus for use with anaerobic cultures. The resulting genome-scale data sets for anaerobic and aerobic growth were validated by comparison to previous small-scale studies comparing growth of *E. coli* under the same conditions. The metabolomics data were then integrated with the *E. coli* genome-scale metabolic model (GEM) via a sensitivity analysis that utilized reaction thermodynamics to reconcile simulated growth rates and reaction directionalities. This analysis highlighted several optimal network usage inconsistencies, including the incorrect use of the beta-oxidation pathway for synthesis of fatty acids. This analysis also identified

enzyme promiscuity for the *pykA* gene, that is critical for anaerobic growth, and which has not been previously incorporated into metabolic models of *E coli*.

Introduction

The physiological and biochemical differences between anaerobic and aerobic growth in bacteria have implications for industrial biotechnology and many areas of human health. The importance of this shift is highlighted by the observation that the transcription factors orchestrating this shift are the most far-reaching in *E. coli.* The impact is a major change in the amount of energy that can be obtained from many commonly used substrates [1-4]. Therefore, much research has been dedicated to understanding this difference, particularly for *E. coli*. Specifically, changes in gene expression [5-7], structural and biochemical differences of preferential anaerobically and aerobically expressed enzymes [8-15], and DNA supercoiling changes [16-18] during aerobicity shifts have been extensively investigated. The flux distribution and ATP maintenance costs have been explored [19], and an integrated omics analysis of aerobic and anaerobic conditions aimed at understanding the effects of nitrate on bacterial growth in the context of global gene regulation has been reported [20]. Even with the vast amount of research that has been dedicated to understanding the physiological and biochemical differences between anaerobic and aerobic organisms, there are still many unanswered questions that could be addressed by examining the differences at the metabolite level.

Advances in metabolomic analytical methods have allowed researchers to accurately probe the metabolome of various microorganisms under various growth conditions [21-23]. By providing a snapshot of the metabolite levels of the cell, metabolomics provides researchers with a powerful tool to investigate the

biochemical status of the cell [24-26]. This information is very useful for the metabolic engineering of microorganisms and the design of new and more specific antibiotics. A prerequisite for meaningful metabolomics data is a robust experimental method. Rapid turnover of metabolites, cell leakage, and residual media are problems that hamper the accurate analysis of microbial metabolism. Various sampling and extraction strategies have been developed and critically evaluated [27, 28] to minimize or circumvent these problems. In addition, rapid sampling apparatuses have been designed to improve the researcher's ability to rapidly and accurately sample cultures [29-31].

Genome-scale modeling techniques have been shown to produce relevant and useful predictions for a diverse set of applications, including metabolic engineering and biological discovery [32]. The development of model-enabled workflows allows for the integration of experimental data types with model-driven analysis to discover new features of biology. In this study, we develop a model-enabled workflow to provide a metabolic snapshot of intracellular metabolite levels for anaerobic and aerobic steady-state growth on glucose minimal media. We describe a modeling-enabled workflow for metabolomics measurements (Fig.1), the analytical and experimental methodology used in this study, the construction of a rapid sampling apparatus, and our QC/QA of the generated metabolomics data. We then provide an analysis of the metabolomics data in the context of the body of research that has been aimed at understanding the differences in anaerobic and aerobic growth. Finally, we integrate the validated metabolomics data set with a genome-scale model of metabolism to examine its

thermodynamic consistency with the current knowledge contained in the metabolic reconstruction.

Material and Methods

Chemicals, unlabeled standards, ^{13}C-labeled standards, and calibration curves

Water, methanol, and acetonitrile were LC-MS grade (Honeywell Burdick & Jackson®). Standards and additives were purchased through Sigma Aldrich or Santa Cruz Biotechnology, Inc. at the highest purity available. Stock solutions were prepared in LC-MS grade water. Compounds with a phosphate group or CoA moiety were titrated to neutral pH with sodium hydroxide to minimize degradation due to spontaneous acid hydrolysis prior to storage at -80 °C. Uniformly labeled ^{13}C-standards were made via metabolic labeling by growing *E. coli* in uniformly labeled Glucose M9 minimal media in aerated shake flasks. All extracts were pooled, well mixed, aliquoted, and stored in the -80 °C for use as internal standards. Calibration mixes of standards were split across several mixes, aliquoted, and lyophilized to dryness. Twelve concentration ranges were obtained by adding the appropriate dilution of water and uniformly labeled standards.

Chromatography, mass spectrometry, and data acquisition

A Synergi™ 2.5 µm Hydro-RP 100 Å LC Column 100 x 2 mm (Phenomenex) with an UFLC XR HPLC (Shimadzu) was used for chromatographic separation, using the gradient profile and mobile phase compositions as described in Lu, et al. [22], with minimal adaptations. The autosampler temperature was 10°C and the injection volume was 10 uL with full loop injection.

An AB SCIEX Qtrap® 5500 mass spectrometer (AB SCIEX) operated in negative mode with multiple reaction monitoring was used for detection and quantification. Electrospray ionization parameters were optimized for 0.2mL/min flow rate, and are as follows: electrospray voltage of -4500 V, temperature of 500 °C, curtain gas of 30, CAD gas of 12, and gas 1 and 2 of 30 and 30psi, respectively. Analyzer parameters were optimized for each compound using manual tuning.

Samples were acquired using the Analyst® 1.6.1 acquisition software and Scheduled MRM™ Algorithm (AB SCIEX). Integration was performed using MultiQuant™ 2.1.1 (AB SCIEX). Compounds were quantified if at least four consecutive points on the calibration curve were found to have a linear coefficient of determination greater than 0.99. The lower limit of quantitation was a signal to noise ratio greater than 40 as calculated by the MultiQuant™ MQ4 integration algorithm, and peak height greater than 1,000 ion counts. For compounds that were incompletely labeled due to the incorporation of exogenous carbon dioxide, calibration curves were truncated at the lowest linear point that was above the signal intensity found in the blank and had minimal bias (i.e., did not regress to zero). Isotope dilution mass spectrometry [33] (IDMS) with metabolically labeled *E. coli* extracts was employed for quantitation. Calibrators and quality controls were included in each batch. In addition, carry over was also monitored with water blanks.

Rapid sampling apparatus construction and operation

The basic premise of the apparatus was taken from the work of Aragon et al, 2006 [34]. The sampling apparatus was designed to allow for rapid sampling and quenching of anaerobic or aerobic batch cultures in triplicate. A schematic of the sampler is shown in Figure 3. A custom microcontroller was designed to control the open time of the solenoid valves, the open time of the pinch valves, the delay between the closing of the solenoid valve and the opening of the pinch valve, and the delay between samples. The microcontroller was fabricated by Brian Millier (Computer Interfaces Consultant). All materials used to fabricate the sampler and a more detailed description of its construction can be found in the Supplementary Information.

Strains, media, and growth conditions

E. coli K12 MG1655 (ATCC 700926), obtained from the American Type Culture Collection (Manassas, VA), was grown under anaerobic and aerobic conditions, as described below, in 4 g/L glucose M9 minimal media [35] with trace elements [36]. Anaerobic cultures were inoculated in an anaerobic chamber (COY; 37 °C; 10% CO2, balance N2) from overnight pre-cultures (starting OD600 ~0.05). Physiological measurements for anaerobic cultures were conducted in the rapid sampling apparatus and anaerobic chamber for comparison. Metabolomic measurements for anaerobic cultures were conducted in the rapid sampling apparatus. Aerobic cultures were inoculated in a fume hood from overnight pre-cultures grown in an incubator (starting OD600 ~0.05). Physiological measurements for aerobic cultures were conducted in the rapid sampling apparatus and water bath only for comparison. Metabolomic

measurements for aerobic cultures were conducted in the rapid sampling apparatus. The water bath was maintained at 37 °C and aerated at 700 RPM.

Physiological measurements for rapid sampler testing

Physiological measurements for culture density were measured at 600 nm absorbance with a spectrophotometer and correlated to cell biomass. Samples to determine substrate uptake and secretion were filtered through a 0.22 µm filter(PVDF, Millipore) and measured using refractive index (RI) detection by HPLC (Waters, MA) with a Bio-Rad Aminex HPX87-H ion exclusion column (injection volume, 10 ul) and 5 mM H2SO4 as the mobile phase (0.5 ml/min, 45°C). Growth, uptake, and secretion rates were calculated from a minimum of four steady-state time-points. Final yield was determined from the amount of starting carbon source and amount of end fermentation products once all of the carbon source was depleted according to the equation $Yp_{/s} = \sum_{i=1}^{n} \left(\frac{\mathrm{C_{product,i} * P_{product,i}}}{\mathrm{C_{substrate} * U_{substrate}}} \right)$ for *n* measured products, where $Yp_{/s}$ is the yield (mmol/mmol), $\mathrm{C_{product,i}}$ is the number of carbons for the ith product, $\mathrm{P_{product,i}}$ is the amount of product secreted (mmol) for the ith product, $\mathrm{C_{substrate}}$ is the number of carbons for the substrate (i.e., glucose), and $\mathrm{U_{substrate}}$ is the amount of substrate consumed (mmol). pH was also monitored for anaerobic cultures.

Evaluation of separation and extraction methods

The separation methods tested involved centrifugation of the culture broth or by application of the differential method [37]. The differential method as described was modified to omit the use of an intermediate quenching solution for

the whole broth and filtrate sample. Samples were instead directly extracted in organic solvent. The extraction solvents tested were either cold, aqueous 80/20 methanol/water (by %vol.) (80/20 MeOH/H2O) [38] at -80 °C or cold, aqueous 40/40/20 acetonitrile with 0.1% formic acid/methanol/water (by %vol.) (40/40/20 ACN/MeOH/H2O) at -20 °C (based on the method described in Rabinowitz et al, 2007 [38]). Internal standards were added to the first extraction solvent, and omitted in subsequent extractions. A detailed description of the modified protocol can be found in the supplemental information.

Comparison of the different sampling and extraction methods was based on 1) total log fold change of normalized endogenous feature signal intensity divided by isotopically labeled feature signal intensity for a representative group of compounds, and 2) log fold percent change in energy charge ratio $\frac{\mathrm{ATP}+{}^{ADP}/_{2}}{ATP+ADP+AMP}$ compared to that of the designated control (i.e., separation by centrifugation and extraction with 80/20 methanol/water (by %vol.)).

Metabolomics of anaerobic and aerobic batch cultures

Anaerobic and aerobic cultures were sampled at three time points using the rapid sampling apparatus during steady-state growth (0.2±0.003, 0.34±0.005, 0.45±0.006) and (1.21±0.07, 2.13±0.06, 3.38±0.04) (OD600, mean±SD), respectively, as described previously using the differential method to separate intracellular metabolites along with the 40/40/20 ACN/MeOH/H2O extraction solvent. The volume of the first extraction was always four times that of the sample volume, and 0.2 mL for the two subsequent extractions, with internal

standards added to the first extraction solution. Reconstitution volumes and the amount of labeled biomass added to the extraction solvent were such that each sample contained 1 gDW/L of biomass and 1 gDW/L of metabolically labeled biomass to account for the differences in biomass between anaerobic and aerobic cultures. The concentration of each sample was normalized to culture density, and the intracellular volume for *E. coli* used to determine the absolute concentrations was taken from Volkmer, et al. 2011[39]. Concentrations from 3 biological replicates and 1 technical replicate were used to determine the average and variance for broth and filtrate samples at each time point. The differential method was applied to those samples whose average filtrate concentration was found to be 80 percent or less than that of the average total broth to determine the average internal concentrations. The internal variance was calculated by adding the variances of the broth and filtrate samples. The final reported data are the averages of the average and variance of the three steady-state time-points.

In silico analysis, modeling, and statistical analysis

All metabolic modeling was performed using the *i*JO1366 genome-scale metabolic network of *E. coli* [40]. All simulations were run using C++ in the Microsoft Visual Studios 10.0 environment with the IBM® ILOG® CPLEX® Optimization Studio v12.3 (IBM®) as the linear program solver. Substrate uptake rates, *in silico* media compositions, and gene deletion simulations were done as described in Orth, et. al. 2011 [40]. Metabolite essentiality [41], flux sum analysis [41], and parsimonious flux balance analysis (pFBA) [42] were performed as described

(see supporting information for a detailed description). Thermodynamic feasibility analysis was performed based on the equations described in Henry et al., 2007 [43] and method described by Zamboni et al., 2008 [44] (see supporting information for a detailed description). Statistical analysis and plots were done using Excel® and Matlab®. P-values were calculated using a Student's t-test. The reaction maps for Figure 6 were generated using Simpheny® (Genomatica) and data was mapped to the figure using in-house scripts.

Results and Discussion

Model-driven analysis of *E. coli's* metabolic pathways to determine target compounds

Given the time and cost that it takes to develop a targeted analytical method, the question of which compounds to measure in a biological system is of paramount importance as the choice of compounds completely shapes the downstream method development. For this purpose, the *E. coli* metabolic model *i*JO1366 was employed to generate a list of target compounds that are important for overall network function. Metabolite usage in the model was calculated by summing metabolic fluxes that utilize a metabolite under a given set of environmental and genetic conditions. While computational metabolite usage is not a true measure of a metabolite's physiological level, the network usage of a compound provides a proxy to identify compounds that are highly utilized and/or change between different genetic and environmental conditions. Following this logic, network usage of the compounds in *E. coli* was simulated and scaled to total network usage of all compounds (Fig. 2). Compounds between each condition that were found to have a high flux and/or high change in flux compared to the control were pooled and rank ordered in order to identify a reasonable number of compounds that could be manually curated (see Methods). The top 15 compounds from this analysis included the adenosine nucleotides, the nicotinamide adenine dinucleotides, acetyl-CoA, coenzyme A, L-glutamate, and phosphoenolpyruvate (PEP), indicating that these metabolites are important to overall network function. Next, literature sources on bacterial

metabolomics, biochemistry textbooks, and the results from metabolite essentiality analysis were used to guide the manual curation step. This step involved manually checking the pool of compounds to give priority to those that are essential for growth, known to have an important role in bacterial physiology, and may not have been identified by the *in silico* model. For instance, 2-oxobutanoate and L-threonine, which were not within the top 100 compounds identified in the previous analysis, were found to be essential for growth and given a higher priority on our list. Similarly, the glutathiones were given increased priority for their role in regulating free radicals. In total, 250 compounds were identified that were desirable to measure in *E. coli*. Based on compound availability and instrumentation (see next section), 126 compounds from this list were optimized for detection.

Development of a quantitative metabolomics method

Analytical method to quantify target compounds

An analytical method was chosen to optimize coverage and accuracy for the measurement of the targeted list of compounds. The target list was dominated by polar, charged species that are involved in central carbohydrate metabolism, amino acid metabolism, nucleotide metabolism, and redox balance. Consequently, an ion-pairing method [22] was chosen for its ability to separate polar species, and an AB SCIEX 5500 Qtrap was chosen for its high sensitivity for quantitation. The platform was manually optimized for detection of each compound (see Methods). All transitions and retention times are provided in Supplemental Table 1.

Experimental method to accurately measure the metabolite levels of anaerobic batch cultures

Considerations for metabolomics sampling include the fast turnover of many intracellular compounds, cell leakage, and residual media that can hamper accurate metabolomics analysis [28]. Furthermore, a methodology suitable to study the range of growth conditions under which *E. coli* can thrive—including both anaerobic and aerobic batch culture—is needed. Given these needs, a rapid sampling technology was developed and various sampling and extraction methods were explored to provide an accurate snapshot of intracellular metabolism.

Rapid sampling apparatus

A first consideration was overcoming the technical challenges involved in sampling anaerobic batch cultures for metabolomics. One accepted approach for anoxic conditions is the use of an anaerobic chamber (see Methods). However, the use of organic solvents for quenching contaminates, the atmospheric composition of the chamber, and the overall design hampers fast sample handling if increased throughput is desired. Therefore, a rapid sampling apparatus based on the previous work of Aragon et al, 2006 [34] that could rapidly sample cultures for biological triplicate was constructed to overcome these shortcomings (Fig. 3). The apparatus was first tested to ensure that it could sample accurately and reproducibly (Supplemental Fig. 2) by sampling cultures of different volumes containing water and by measuring the volume sampled gravimetrically. For the pressures used in this work, the effect of culture volume

on sampling volume was negligible. The generated sampling apparatus was found to eject a sample from the culture vessel into the extraction solution in less than a second. The limiting factor in sample frequency was the speed at which the operator could change the purge vessel (i.e., the previous sample vessel to the next sample vessel). For one operator, it took approximately 10 seconds, and for multiple operators, it took approximately 3 seconds to change the sample vessel.

The developed rapid sampling apparatus was validated to ensure that it could reproduce culture conditions within the culturing vessel as compared to a laboratory 'gold standard' (i.e., a water bath for aerobic cultures and a temperature controlled anoxic chamber for anaerobic cultures). The results from this analysis are given in Table 1 and Supplemental Figure 1. Comparisons between the sampling apparatus and our 'gold standard' culture conditions were based on growth rate, substrate uptake/secretion rates during steady-state, and overall carbon yield at the conclusion of the fermentation. For all three criteria, the values obtained between the sampling apparatus and our culture conditions were found to be consistent (the mean growth rates, glucose uptake rates, and carbon yields deviated less than 3%, 4%, and 4% from the 'gold standard' values, respectively).

Sampling and extraction methodology

Modified versions of several published methods were compared to select a sampling and extraction protocol for our experimental conditions. Several sampling methods [24, 37, 45] and extraction solvents [27, 38] were initially considered.

To differentiate internal from external compounds, separation by centrifugation or application of the differential method [37] were tested. To precipitate proteins and extract metabolites, cold, organic extraction solvents consisting of 80/20 methanol/water and 40/40/20 acetonitrile + 0.1% formic acid/methanol/water (by %vol.) were also tested. In addition, the classic Bligh-Dyer cholorform/methanol extraction method [46] was initially tested, but an unacceptable degradation of phosphorylated compounds was found during initial experiments (data not shown), and the method was not pursued further.

Each combination of separation and extraction was tested and assessed for steady-state, aerobic cultures using the rapid sampling apparatus (Fig. 4). Briefly, the criteria to judge the different approaches were total log fold signal change and log fold percent change in energy charge ratio. A large difference in fold signal intensity (particularly for the phosphorylated compounds) was found between centrifuged samples and samples separated by the differential method. For compounds with slow turnover times (e.g., glutamate), differences were smaller. A large difference in log fold percent change in energy charge ratio between sampling methods was also found. It should be noted that on average, the calculated energy charge ratio for the optimized sampling and extraction method was 0.95 and 0.94 for anaerobic and aerobic conditions, respectively. This is within the physiological range of 0.8 – 1.0 [37]. The findings above indicate that a sampling method that provides near instantaneous quenching of metabolism is needed for compounds with fast turnover times (e.g., ATP),. Therefore, it was decided to use direct extraction and application of the

differential method, and extraction by the acetonitrile-based solvent due to its higher log fold energy charge ratio.

While providing the fastest possible arrest of cellular metabolism for liquid cultures, one drawback to the use of the method described here is ion suppression and interference from media components due to increased sample matrix. Ion suppression from sample matrix, particularly in the early stages of the LC gradient, was evident. Extracellular interference has been attributed to cell lysis, as well as specific and non-specific membrane transporters [28]. Due to the large volume of the medium, even small amounts of extracellular components can greatly distort the analysis if not properly accounted for [28]. We therefore limited our analysis to only those components that were found to have an extracellular concentration of less than 80% of that contained in the whole broth. This limited the number of components for which we could provide quantitative data out of the total that we were monitoring in each sample. For example, cAMP was found in equal amounts in the whole broth and filtrate samples, and could not be included in this study.

Interpretation and validation of the steady-state anaerobic and aerobic metabolomes of *E. coli*

Of the total of 50 metabolites measured across both conditions which passed the quality control specifications (see Methods), 38 were measured under anaerobic conditions, 49 were measured under aerobic conditions, and 37 were measured in both conditions (Supplemental Table 2). Of the 37 shared metabolites, 15 were significantly different between the conditions (p-value <

0.01), of which 11 had a fold change greater than 2 between the two conditions (Fig. 5). In order to understand these differences, a literature comparison was performed. Enzymatic, expression, and various other data types applied to investigate various aspects of anaerobic and aerobic physiology over the last few decades were integrated to validate findings and provide contextual support for several of our other findings regarding differences between anaerobic and aerobic growth (Table 3).

NADH, NADPH, Reduced glutathione, and L-glutamine were found to have large and statistically significant differences between anaerobic and aerobic growth. The increased levels of NADH found in anaerobic cultures are indicative of an inactive electron transport chain [1, 47]. The reduced levels of NADPH and Reduced glutathione found in anaerobic cultures provide evidence for increased flux through glucose-6-phosphate dehydrogenase and a cellular response to oxygen radicals generated from an active electron transport chain during aerobic growth [48-50]. The decreased levels of glutamine (and glutamate) were found under anaerobic conditions compared to aerobic conditions, which indicate a more acidic cytosol and decreased flux towards glutamate precursors under anaerobic growth [51-54]. A more detailed discussion is provided in the supplemental information.

Elevated levels of NTPs were also found in anaerobic cultures. One possible explanation for this is that NTP generation is more closely linked to glycolysis under anaerobic conditions than under aerobic conditions. NDP kinase is the primary source for NTPs under aerobic conditions. NDP mutants

are able to synthesize NTPs through adenylate kinase under aerobic growth. However, adenylate kinase shows broad, but preferential, specificity [55] leading to imbalances in the NTP, which has been found to lead to increased mutagenesis [56]. In contrast to aerobic growth, increased mutagenesis has not been found in NDP kinase mutants under anaerobic growth [56], suggesting that another enzyme provides the NTP synthesizing machinery. Current evidence suggests that *pykA* is the dominant nucleotide kinase under anaerobic conditions [10]. *PykA* is positively regulated by FNR [57] and shows broad and largely non-preferential substrate specificity [10]. Interestingly, our metabolomics data indicate that the NDP kinase function of converting UDP to UTP at the expense of ATP is thermodynamically infeasible under anaerobic growth. Removal of this reaction from the network yields no growth under anaerobic conditions. However, upon addition of the C-, G-, U-, and dT-TP kinase activity to pyruvate kinase, maximal growth was predicted. This was found to hold even when all reactions catalyzed by NDP kinase were removed from the network. This provides modeling and metabolomics data to support the hypothesis that the NTP pool is largely generated through glycolysis via pyruvate kinase.

Integrating measurements with modeling: assessing thermodynamics at the genome-scale

Analysis of thermodynamics at the genomic-scale allows for the assessment of the thermodynamic feasibility of a metabolomics data set. Therefore, a thermodynamic feasibility analysis based on the equations described in Henry et al., 2007 [43] and method described by Zamboni et al., 2008

[44] was performed. In addition to adopting these analyses, additional work was performed to explore the cost of constraining reaction fluxes on the network, and to explore the overall feasibility of catabolic and anabolic pathways. The analysis of the metabolomics data presented here resulted in the identification of one thermodynamically infeasible reaction and two reactions that were thermodynamically inconsistent with optimal growth. Specifically, the infeasible reaction was found under anaerobic conditions, and the two thermodynamically inconsistent reactions were found under both anaerobic and aerobic conditions. The results are presented in Table 2 and described further below.

The nucleoside-diphosphate (NDP) kinase (ndk) reaction which converts UDP to UTP at the cost of ATP was found to be thermodynamically infeasible when analyzing the anaerobic data set with the iJO1366 *E. coli* model (Table 2). This reaction was investigated further. There is evidence to suggest that under anaerobic growth, the nucleotide triphosphate (NTP) pool is maintained by pyruvate kinase [10]. Currently, in the reconstructed model, both the pykA and pykF gene products were assigned to only one reaction: adp + h + pep --> atp + pyr, EC: 2.7.1.40. There is also evidence that pyruvate kinase can transfer phosphate from PEP to other NTPs (CDP, GDP, UDP, or dTDP) to form pyruvate and the respective triphosphate by the predominantly anaerobically expressed pyruvate kinase isozyme, *pykA* [10]. As such, these reactions were temporarily added to the model to mimic the broad kinase activity of the *pykA* enzyme found *in vitro*. Upon changing the NDP kinase reaction to account for enzyme promiscuity, no penalty in growth rate was predicted and the infeasibility

disappeared. This indicated that NDP kinase is not essential for growth under anaerobic conditions, and that the thermodynamic infeasibility of the NDP kinase reaction was not a consequence of the metabolomics data, but an error in the metabolic model. Therefore, this integration of metabolomic data with the model points to an oversight in promiscuity for an enzyme [58] that has been present for four major revisions of the widely-used *E. coli* metabolic reconstruction

A thermodynamically inconsistent reaction identified in the integrated analysis corresponded to acetyl-CoA C-acetyltransferase (atoB) for both the anaerobic and aerobic growth conditions (Table 2). Acetyl-CoA C-acetyltransferase is a component of the beta-oxidation pathway that catalyzes the conversion of acetoacetyl-CoA and coenzyme A to acetyl-CoA, which is required for growth on acetoacetate [59]. Optimal growth is predicted to use the beta-oxidation pathway, which does not require the reducing power of NADPH to synthesize lipids. Constraining the direction of flux through this reaction resulted in a 2.9% and 1.1% reduction in growth under anaerobic and aerobic conditions, respectively, due to the cost of utilizing NADPH for lipid synthesis. Under normal physiological conditions, lipid synthesis proceeds first by carboxylation of acetyl-CoA to malonyl-CoA, and then elongation continues using the reducing power of NADPH [60]. The energetic advantage of using the reversed beta-oxidation pathway for the production of fatty acids was recently demonstrated for metabolically engineered *E. coli* [61].

The reaction corresponding to the phosphoglycerate dehydrogenase (serA) gene product was found to be thermodynamically inconsistent for both the

anaerobic and aerobic growth conditions (Table 2). Phosphoglycerate dehydrogenase catalyzes the first committed step to serine synthesis [62]. This one reaction has been shown to be thermodynamically unfavorable along with a high Km for 3-phosphoglycerate [62]. The entire pathway (which *serA* belongs to) is instead driven in the forward direction by the subsequent reaction catalyzed by phosphoserine aminotransferase (serC) [62]. To evaluate this finding with the model, a pathway feasibility analysis for serine biosynthesis was conducted. It was reasoned that the thermodynamics of a pathway, as compared to a single reaction, depends to a greater degree on physiologically relevant ratios of currency metabolites (e.g. ATP/ADP, NAD(P)/NAD(P)H, L-glu/L-gln). For the serine biosynthetic pathway, the *in vivo* thermodynamics were found to be consistent with serine biosynthesis. As another thermodynamic check of the data, we expanded this approach to conduct a similar pathway-centric thermodynamic analysis for glycolysis, the oxidative pentose phosphate pathway, aspartate, threonine, tryptophan, tyrosine, phenylalanine, and arginine biosynthesis, and *de novo* purine and pyrimidine biosynthesis (Supplemental Table 3). Under all conditions the pathways were found to be thermodynamically feasible. Although disagreements and errors often point to new biology in integrated modeling analyses (as seen above), these consistent findings give confidence in the generated data sets and models as they are in agreement with previously reported findings.

Metabolomics data coverage:

The number of reactions that could be assigned a thermodynamic change in free energy from our measured anaerobic and aerobic metabolomics data was compared to the number of values that could be theoretically assigned by choosing a set of metabolites at random. This was done in order to assess the reaction coverage achieved by the model-driven approach to select metabolites in this study and was examined by reaction subsystem. For 50 randomly chosen metabolites (i.e., the combined number of unique metabolites measured for anaerobic and aerobic conditions) repeatedly selected 10,000 times from the pool of 500 metabolites which were found to be used most in the cell (Fig.2, Step 2), there was less than a 4% chance of achieving sufficient metabolite concentration coverage to make a single reaction call for any of the 38 subsystems in the reconstruction. When analyzing the set of metabolites chosen for measurement in this work, it was possible to make at least one reaction call for 22 subsystems under either condition (22 and 18 for aerobic and anaerobic, respectively, see Supplementary Figure 3). The mean and median numbers of reaction values estimated per subsystem were 3.2 and 2, respectively, for the aerobic set, and 2.7 and 1.5, respectively, for the anaerobic set. This reaffirms the benefit of an *a priori* targeted analysis to identify compounds to maximize network coverage from a limited set of measurements.

Conclusion

Model-driven analyses provide a context in which omics data types can be mapped onto genome-scale models to identify biological differences between conditions of interest and further improve the knowledge collected in GEMs. This approach was applied to analyze data generated by LC-MS/MS from anaerobic and aerobic steady-state cultures sampled using an atmosphere-controlled sampling apparatus. In particular, thermodynamic analysis was used to calculate *in vivo* reaction free energies from quantitative metabolomics data between these two conditions in the context of the GEM. This analysis highlighted inconsistencies between modeling objectives and *in vivo* function. An incorrect use of the beta-oxidation pathway for synthesis of fatty acids was identified, as well as an incorrect use of the dNTP synthesizing machinery for DNA replication under anaerobic growth. This latter finding provided insight on enzyme promiscuity that is vital for anaerobic growth. Thus, a combination of LC-MS-generated metabolomics and modeling provide further support for the importance of enzyme promiscuity for normal physiological growth.

Model-driven analyses can also provide a context for method development. The predicted network usage under various growth conditions of interest was used to identify a target list of compounds for downstream metabolomics method development that we hypothesized would provide insight into bacterial biochemistry. When compared to a set of randomly selected metabolites in the cell, the set chosen here facilitated the prediction of far more reaction and pathway thermodynamic values versus a list selected at random.

Therefore, the model-driven approach provided a means to get the most information from a limited set of resources (i.e., the number of metabolite measurements one can measure in one sample injection). Furthermore, the physiochemical properties of the target list guided the choice of analytical platform and experimental methodology. Due to the general chemical liability of the majority of compounds in our target list and the growth conditions of interest, specific steps were necessary for an appropriate experimental method development. Among others, this included the construction of a sampling apparatus for fast and accurate sampling of anoxic, liquid cultures.

This study demonstrates how GEMs can be incorporated into a workflow that includes model-enabled method development and model-driven data analyses. The approach can be extended to different organisms where metabolic reconstructions are readily available and to additional growth conditions of industrial or environmental interest where a functional comparative analysis at the metabolic and reaction levels is desirable.

Acknowledgments

Chapter 3, in full, is a reformatted reprint of “A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent.” McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, Feist AM. Biotechnol Bioeng. 2014 Apr;111(4):803-15. doi: 10.1002/bit.25133. The dissertation/thesis author was the primary investigator and author of the paper.

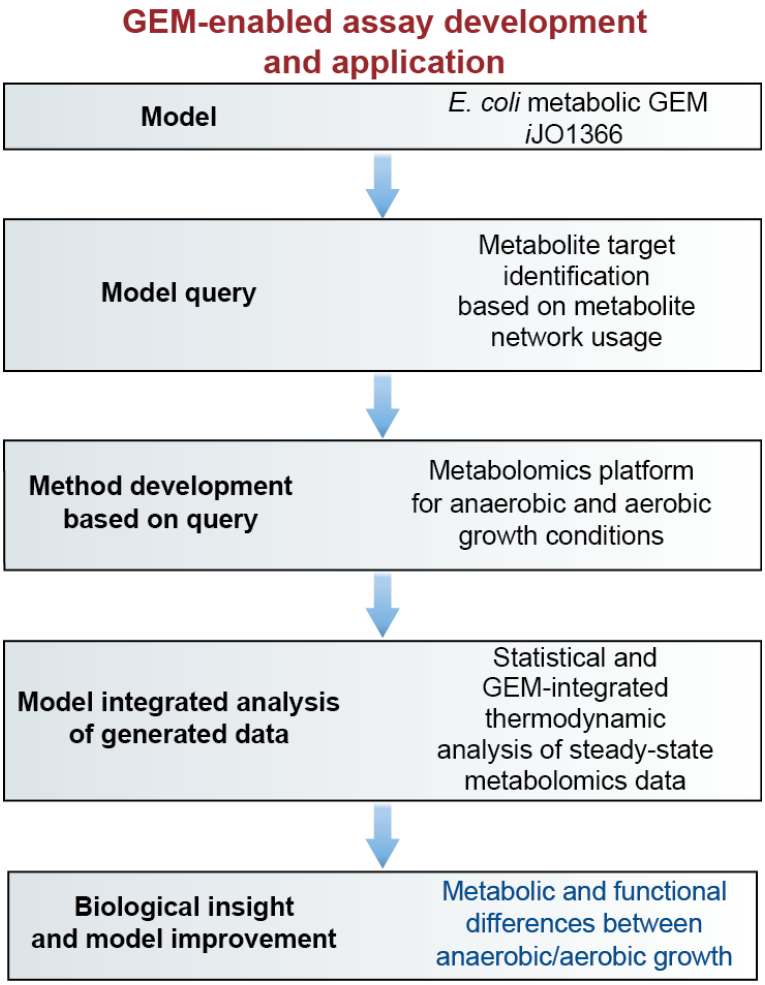

Figure 3.1: The GEM-enabled workflow utilized for the development of a metabolomics assay to study the biochemical differences between anaerobic and aerobic growth of *E. coli* K-12 MG1655.

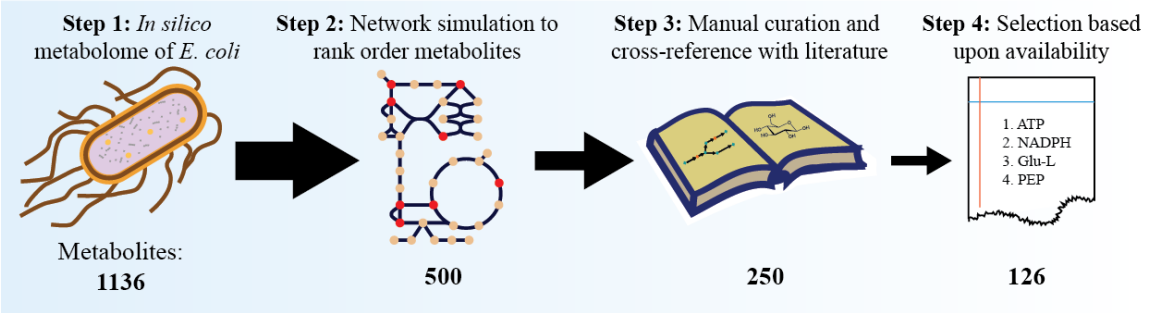

Figure 3.2: Identification of target compounds for subsequent method development. The process was initiated by considering all of the unique metabolites in the latest genome-scale metabolic reconstruction of *E. coli* (step 1). Next, simulations were performed under a wide range of potential genetic and environmental conditions to determine and rank order metabolites that are the most highly used and/or whose usage changes the most between conditions (step 2). Metabolite essentiality was also performed for anaerobic and aerobic conditions. A reduced list of target compounds was generated by manually selecting compounds based on their rank, their essentiality to network function, and known biochemical or physiological importance from literature sources (step 3). Finally, the list was screened based on availability for purchase (step 4).

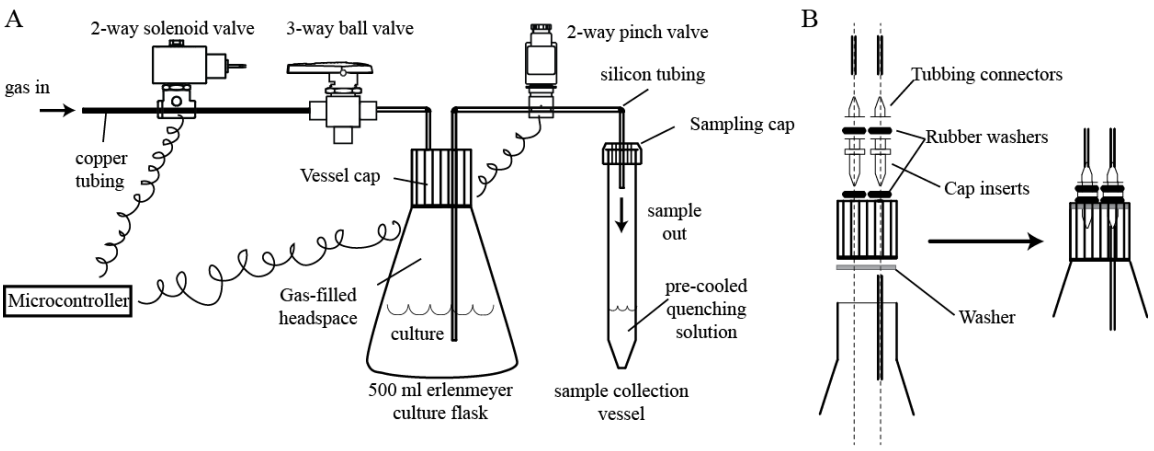

Figure 3.3: Schematic of the anaerobic rapid sampling apparatus. A) Overview of the device, and B) Detailed view of the culture vessel caps.

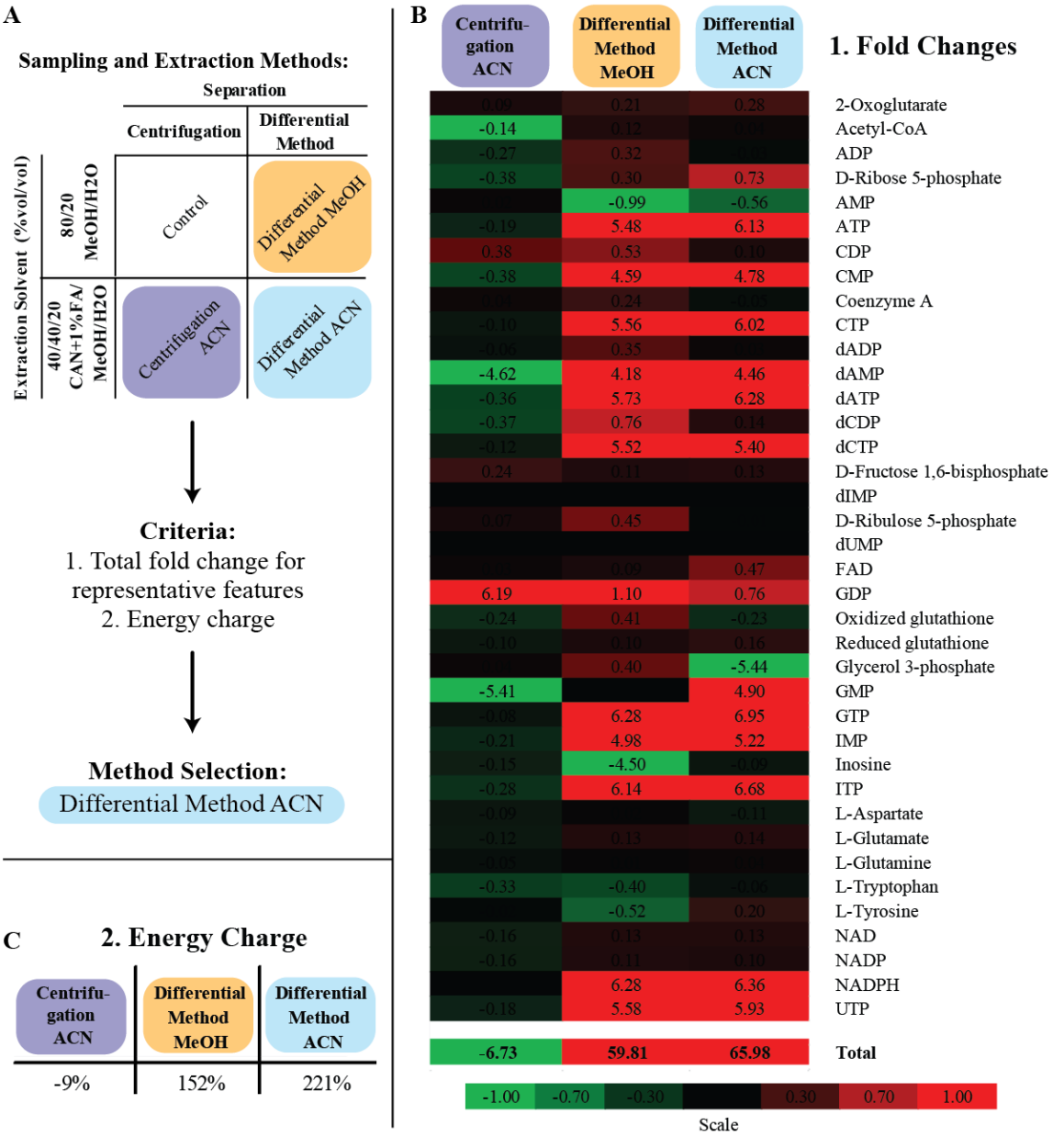

Figure 3.4: Separation and extraction solvent method comparison. A) Table of combinations of separation and extraction solvent method tested, and workflow to decide upon an appropriate separation and extraction solvent method. B) Heat map of unlabeled to labeled signal ratios for compounds used for method development. C) Energy change between different separation and extraction methods.

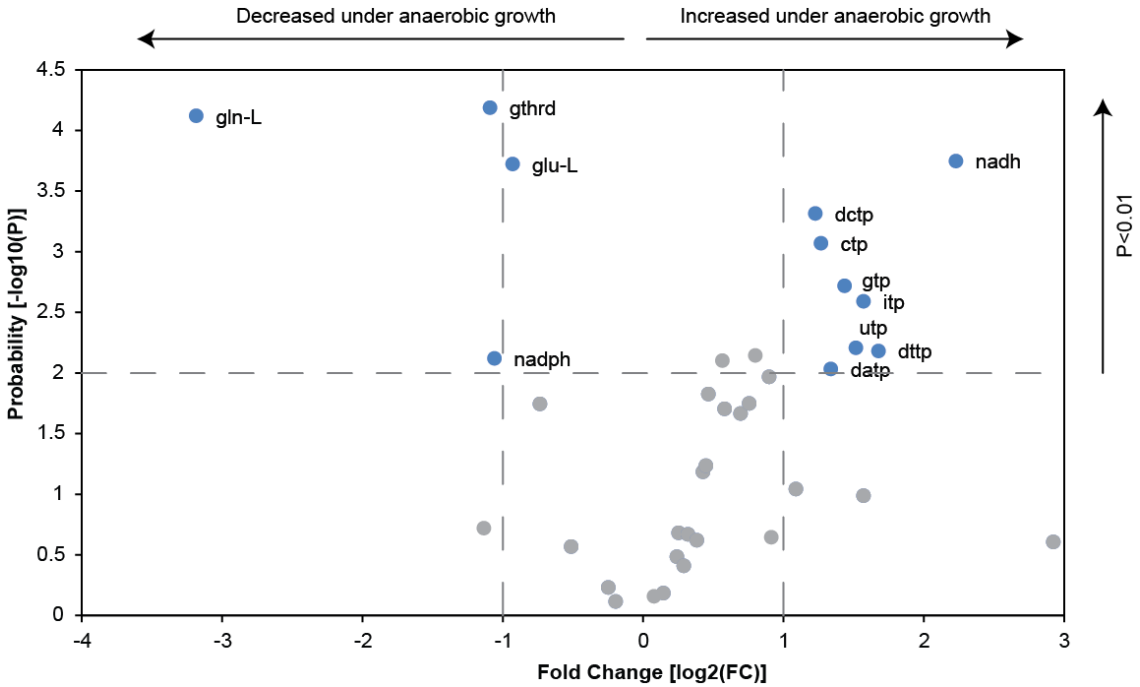

Figure 3.5: A comparison of anaerobic versus aerobic metabolism. Compounds with a significant change (P<0.01) are highlighted in blue. X-axis refers to log 2 fold changes in metabolite concentrations between anaerobic and aerobic steady-state cultures. Y-axis refers to – log 10 p-values in metabolite concentrations between anaerobic and aerobic steady-state cultures.

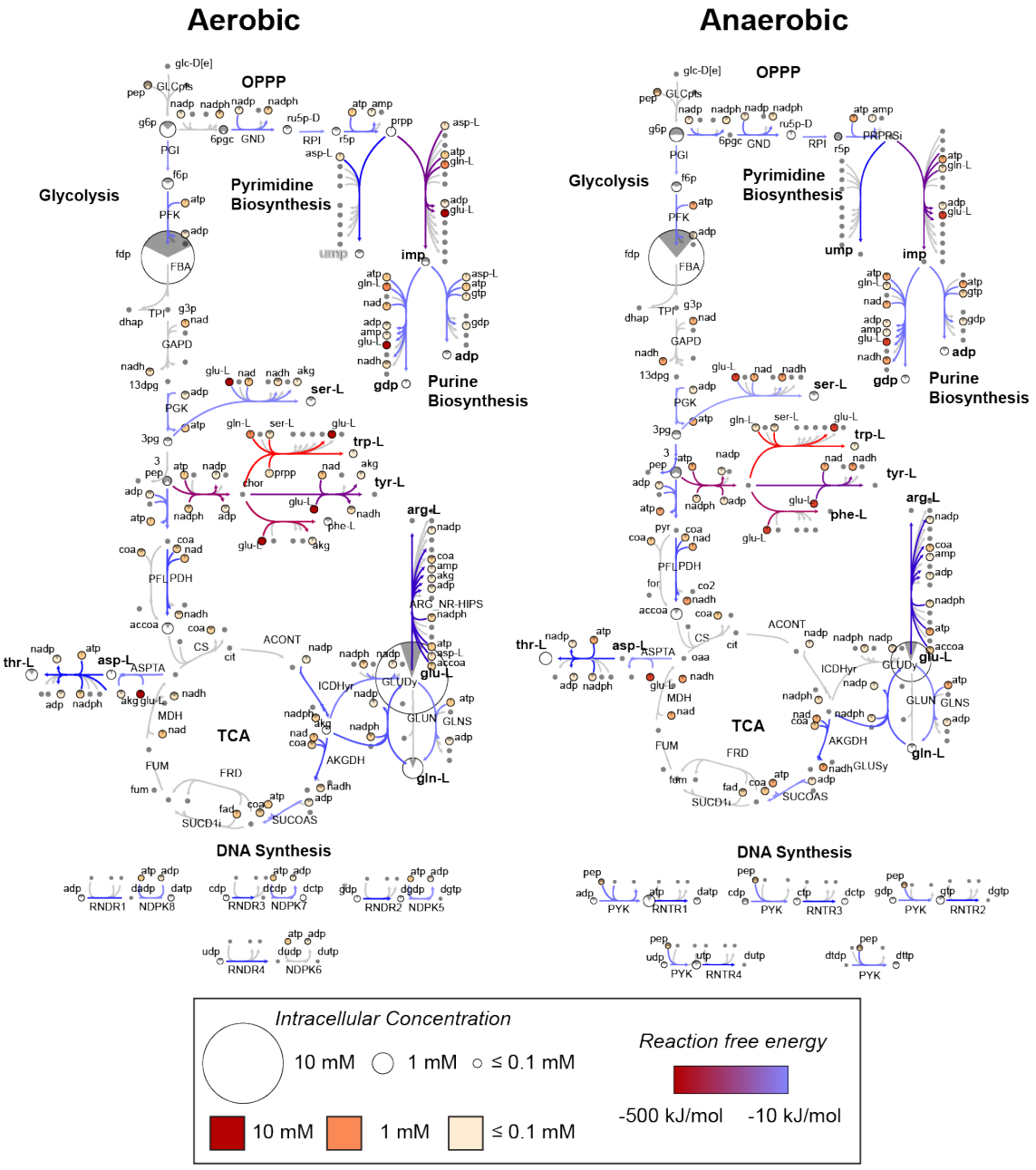

Figure 3.6: A visual integration of metabolomics data with the metabolic model of *E. coli*. Intracellular concentrations of measured metabolites (aerobic on the left and anaerobic on the right) are mapped onto pathways from the model and concentrations are used to scale metabolites (nodes) by radius. Intracellular concentrations for cofactor components are scaled by color. The gray insert on each metabolite is scaled such that the percent of the area filled is the percent coefficient of variance. Reaction links are scaled according to reaction free energy as calculated from the intracellular concentrations. Gray metabolites and flux arrows were not measured or calculated.

Table 3.1: Physiological comparison between anaerobic and aerobic cultures taken from the rapid sampling apparatus and those from our traditional culture conditions.

		Growth Rate	Glucose Uptake Rate	Carbon Conversion
		hr-1	mmol*gDW-1*hr-1	Yield ($Y_{p/s}$)
Rapid Sampler	Anaerobic	0.38 ± < 0.01	13.7 ± 1.2	1.00 ± 0.07
	Aerobic	0.59 ± 0.01	7.81 ± 0.27	0.01 ± < 0.01
Control	Anaerobic	0.39 ± < 0.01	13.7 ± 0.4	0.97 ± 0.02
	Aerobic	0.61 ± 0.02	7.54 ± 0.56	0.00 ± < 0.01

Table 3.2: Thermodynamically infeasible reactions. * After removal of reactions for NDP kinase and addition of pyruvate kinase reactions (pykA) to account for broad substrate specificity for U, C, G, and dT-triphosphorylated nucleotides under anaerobic growth.

	Reaction Name	Reaction Formula	GrSTD (kJ/mol)	Grlb (kJ/mol)	Grub (kJ/mol)	% reduction in growth	% synergistic reduction in growth	
Aerobic	acetyl-CoA C-acetyltransferase	2 accoa[c] -> aacoa[c] + coa[c]	26.6	18.1	36.7	1.1	5.9	
	phosphoglycerate dehydrogenase	3pg[c] + nad[c] -> 3php[c] + h[c] + nadh[c]	14.9	8.6	26.7	3.8		
Anaerobic	acetyl-CoA C-acetyltransferase	2 accoa[c] -> aacoa[c] + coa[c]	26.6	18.2	36.1	2.9	5.8	5.8*
	phosphoglycerate dehydrogenase	3pg[c] + nad[c] -> 3php[c] + h[c] + nadh[c]	14.9	10.9	27.3	3		
	nucleoside-diphosphate kinase (ATP:UDP)	atp[c] + udp[c] -> adp[c] + utp[c]	0	0.3	0.9	0*		

Table 3.3: Biological implications of differences found in the metabolite levels of anaerobic cultures compared to aerobic cultures.

Finding	Biological Implication	Literature Support
Increased NADH	An inactive electron transport chain, a more reduced quinine pool, and the expression of the ArcAB two component system.	[1, 47, 63]
Decreased NADPH	Reduced flux through the oxidative PPP, through isocitrate dehydrogenase, [and through the membrane-bound transhydrogenase (PntAB).	[19, 64]
Decreased NADPH	Reduced need for protection against superoxide radicals (Increased expression of SoxRS and OxyR in aerobic cultures)	[47]
Decreased Reduced Glutathione	Reduced NADPH synthesis, and reduced generation of oxygen radicals (Increased expression of SoxRS and OxyR in aerobic cultures)	[16, 47]
Decreased glutamate and glutamine	Increased cytosolic pH (i.e. mild organic acid stress)	[52, 65, 66]
Decreased glutamate and glutamine	Increased cytosolic ionic strength (i.e. mild osmotic stress)	[51, 67]
Decreased glutamate and glutamine	Decreased flux through 2-oxoglutarate, and down regulation of gltA (citrate synthase) by arcA	[53, 54]
Relatively constant or slightly increased ATP/ADP ratio	Increased DNA supercoiling maintained primarily by a reduction in Topoisomerase I activity during steady-state.	[18]
Increased dNTP pool	A reduced number of replication forks.	[68]
Increased dNTP pool	Expression of the anaerobic ribonucleotide reductase (nrdDG), which does not appear to be strictly regulated by feedback from the dNTP pool.	[12, 13, 69]
Increased NTP pool	Increased expression of pykA (coupling of NTP pool formation primarily through glycolysis instead of ribonuceotide diphosphate kinase activity).	[10, 56, 57]
Increased NTP pool	Substrate preference of NTPs instead of NDPs by the anaerobic ribonucleotide reductase.	[9]

References:

1. de Graef, M.R., Alexeeva, S., Snoep, J.L. & Teixeira de Mattos, M.J. The steady-state internal redox state (NADH/NAD) reflects the external redox state and is correlated with catabolic adaptation in Escherichia coli. *J Bacteriol* **181**, 2351-2357 (1999).

2. Rolfe, M.D. Transcript profiling and inference of Escherichia coli K-12 ArcA activity across the range of physiologically relevant oxygen concentrations. *J Biol Chem* **286**, 10147-10154 (2011).

3. Green, J. & Paget, M.S. Bacterial redox sensors. *Nat Rev Microbiol* **2**, 954-966 (2004).

4. Trotter, E.W. Reprogramming of Escherichia coli K-12 metabolism during the initial phase of transition from an anaerobic to a micro-aerobic environment. *PLoS ONE* **6**, e25501 (2011).

5. Salmon, K.A. Global Gene Expression Profiling in Escherichia coli K12: EFFECTS OF OXYGEN AVAILABILITY AND ArcA. *Journal of Biological Chemistry* **280**, 15084-15096 (2005).

6. Shalel-Levanon, S., San, K.Y. & Bennett, G.N. Effect of oxygen, and ArcA and FNR regulators on the expression of genes related to the electron transfer chain and the TCA cycle in Escherichia coli. *Metab Eng* **7**, 364-374 (2005).

7. Tolla, D.A. & Savageau, M.A. Regulation of aerobic-to-anaerobic transitions by the FNR cycle in Escherichia coli. *J Mol Biol* **397**, 893-905 (2010).

8. Walsby, C.J., Ortillo, D., Broderick, W.E., Broderick, J.B. & Hoffman, B.M. An anchoring role for FeS clusters: chelation of the amino acid moiety of S-adenosylmethionine to the unique iron site of the [4Fe-4S] cluster of pyruvate formate-lyase activating enzyme. *J Am Chem Soc* **124**, 11270-11271 (2002).

9. Nelson, D. & Cox, M. Lehninger Principles of Biochemistry, Fourth Edition. (W. H. Freeman, 2004).

10. Saeki, T., Hori, M. & Umezawa, H. Pyruvate kinase of Escherichia coli. Its role in supplying nucleoside triphosphates in cells under anaerobic conditions. *J Biochem* **76**, 631-637 (1974).

11. Ponce, E., Flores, N., Martinez, A., Valle, F. & Bolivar, F. Cloning of the two pyruvate kinase isoenzyme structural genes from Escherichia coli: the relative roles of these enzymes in pyruvate biosynthesis. *J Bacteriol* **177**, 5719-5722 (1995).

12. Tamarit, J. The activating component of the anaerobic ribonucleotide reductase from Escherichia coli. An iron-sulfur center with only three cysteines. *J Biol Chem* **275**, 15669-15675 (2000).

13. Tamarit, J., Mulliez, E., Meier, C., Trautwein, A. & Fontecave, M. The anaerobic ribonucleotide reductase from Escherichia coli. The small protein is an activating enzyme containing a [4fe-4s](2+) center. *J Biol Chem* **274**, 31291-31296 (1999).

14. Gon, S. A novel regulatory mechanism couples deoxyribonucleotide synthesis and DNA replication in Escherichia coli. *EMBO J* **25**, 1137-1147 (2006).

15. Martin, J.E. & Imlay, J.A. The alternative aerobic ribonucleotide reductase of Escherichia coli, NrdEF, is a manganese-dependent enzyme that enables cell replication during periods of iron starvation. *Mol Microbiol* **80**, 319-334 (2011).

16. Bhriain, N.N., Dorman, C.J. & Higgins, C.F. An overlap between osmotic and anaerobic stress responses: a potential role for DNA supercoiling in the coordinate regulation of gene expression. *Molecular Microbiology* **3**, 933-942 (1989).

17. Hsieh, L.S., Burger, R.M. & Drlica, K. Bacterial DNA supercoiling and [ATP]/[ADP]. Changes associated with a transition to anaerobic growth. *J Mol Biol* **219**, 443-450 (1991).

18. Cortassa, S. & Aon, M.A. Altered topoisomerase activities may be involved in the regulation of DNA supercoiling in aerobic-anaerobic transitions in Escherichia coli. *Mol Cell Biochem* **126**, 115-124 (1993).

19. Chen, X., Alonso, A.P., Allen, D.K., Reed, J.L. & Shachar-Hill, Y. Synergy between (13)C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in E. coli. *Metab Eng* **13**, 38-48 (2011).

20. Toya, Y., Nakahigashi, K., Tomita, M. & Shimizu, K. Metabolic regulation analysis of wild-type and arcA mutant Escherichia coli under nitrate conditions using different levels of omics data. *Mol Biosyst* **8**, 2593-2604 (2012).

21. Buscher, J.M., Czernik, D., Ewald, J.C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal Chem* **81**, 2135-2143 (2009).

22. Lu, W. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal Chem* **82**, 3212-3221 (2010).

23. Buescher, J.M., Moco, S., Sauer, U. & Zamboni, N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. *Anal Chem* **82**, 4403-4412 (2010).

24. Bennett, B.D., Yuan, J., Kimball, E.H. & Rabinowitz, J.D. Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. *Nat Protoc* **3**, 1299-1311 (2008).

25. Ishii, N. Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science* **316**, 593-597 (2007).

26. Taymaz-Nikerel, H., van Gulik, W.M. & Heijnen, J.J. Escherichia coli responds with a rapid and large change in growth rate upon a shift from glucose-limited to glucose-excess conditions. *Metab Eng* **13**, 307-318 (2011).

27. Winder, C.L. Global metabolic profiling of Escherichia coli cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. *Anal Chem* **80**, 2939-2948 (2008).

28. Bolten, C.J., Kiefer, P., Letisse, F., Portais, J.C. & Wittmann, C. Sampling for metabolome analysis of microorganisms. *Anal Chem* **79**, 3843-3849 (2007).

29. Hiller, J., Franco-Lara, E., Papaioannou, V. & Weuster-Botz, D. Fast sampling and quenching procedures for microbial metabolic profiling. *Biotechnol Lett* **29**, 1161-1167 (2007).

30. Lange, H.C. Improved rapid sampling for in vivo kinetics of intracellular metabolites in Saccharomyces cerevisiae. *Biotechnol Bioeng* **75**, 406-415 (2001).

31. Schaub, J., Schiesling, C., Reuss, M. & Dauner, M. Integrated sampling procedure for metabolome analysis. *Biotechnol Prog* **22**, 1434-1442 (2006).

32. McCloskey, D., Palsson, B.O. & Feist, A.M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* **9**, 661 (2013).

33. Wu, L. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards. *Anal Biochem* **336**, 164-171 (2005).

34. Aragon, A.D. An automated, pressure-driven sampling device for harvesting from liquid cultures for genomic and biochemical analyses. *J Microbiol Methods* **65**, 357-360 (2006).

35. Sambrook, J., and D. W. Russell Molecular cloning: a laboratory manual, 3rd ed., vol. A2.2. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001).

36. Fong, S.S. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

37. Taymaz-Nikerel, H. Development and application of a differential method for reliable metabolome analysis in Escherichia coli. *Anal Biochem* **386**, 9-19 (2009).

38. Rabinowitz, J.D. & Kimball, E. Acidic acetonitrile for cellular metabolome extraction from Escherichia coli. *Anal Chem* **79**, 6167-6173 (2007).

39. Volkmer, B. & Heinemann, M. Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS ONE* **6**, e23126 (2011).

40. Orth, J.D. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* **7**, 535 (2011).

41. Kim, P.J. Metabolite essentiality elucidates robustness of Escherichia coli metabolism. *Proc Natl Acad Sci U S A* **104**, 13638-13642 (2007).

42. Lewis, N.E. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology* **6**, 390 (2010).

43. Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys J* **92**, 1792-1805 (2007).

44. Zamboni, N., Kummel, A. & Heinemann, M. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9**, 199 (2008).

45. Kleijn, R.J. Metabolic fluxes during strong carbon catabolite repression by malate in Bacillus subtilis. *J Biol Chem* **285**, 1587-1596 (2010).

46. Bligh, E.G. & Dyer, W.J. A rapid method of total lipid extraction and purification. *Can J Biochem Physiol* **37**, 911-917 (1959).

47. Partridge, J.D., Scott, C., Tang, Y., Poole, R.K. & Green, J. Escherichia coli transcriptome dynamics during the transition from anaerobic to aerobic conditions. *J Biol Chem* **281**, 27806-27815 (2006).

48. Imlay, J.A. Cellular defenses against superoxide and hydrogen peroxide. *Annu Rev Biochem* **77**, 755-776 (2008).

49. Giro, M., Carrillo, N. & Krapp, A.R. Glucose-6-phosphate dehydrogenase and ferredoxin-NADP(H) reductase contribute to damage repair during the soxRS response of Escherichia coli. *Microbiology* **152**, 1119-1128 (2006).

50. Krapp, A.R., Humbert, M.V. & Carrillo, N. The soxRS response of Escherichia coli can be induced in the absence of oxidative stress and oxygen by modulation of NADPH content. *Microbiology* **157**, 957-965 (2011).

51. Underwood, S.A., Buszko, M.L., Shanmugam, K.T. & Ingram, L.O. Lack of protective osmolytes limits final cell density and volumetric productivity of ethanologenic Escherichia coli KO11 during xylose fermentation. *Appl Environ Microbiol* **70**, 2734-2740 (2004).

52. Roe, A.J., McLaggan, D., Davidson, I., O'Byrne, C. & Booth, I.R. Perturbation of anion balance during inhibition of growth of Escherichia coli by weak acids. *J Bacteriol* **180**, 767-772 (1998).

53. Underwood, S.A., Buszko, M.L., Shanmugam, K.T. & Ingram, L.O. Flux through citrate synthase limits the growth of ethanologenic Escherichia coli KO11 during xylose fermentation. *Appl Environ Microbiol* **68**, 1071-1081 (2002).

54. Park, S.J., McCabe, J., Turna, J. & Gunsalus, R.P. Regulation of the citrate synthase (gltA) gene of Escherichia coli in response to anaerobiosis and carbon supply: role of the arcA gene product. *J Bacteriol* **176**, 5086-5092 (1994).

55. Lu, Q. & Inouye, M. Adenylate kinase complements nucleoside diphosphate kinase deficiency in nucleotide metabolism. *Proc Natl Acad Sci U S A* **93**, 5720-5725 (1996).

56. Lu, Q., Zhang, X., Almaula, N., Mathews, C.K. & Inouye, M. The gene for nucleoside diphosphate kinase functions as a mutator gene in Escherichia coli. *J Mol Biol* **254**, 337-341 (1995).

57. Shalel-Levanon, S., San, K.Y. & Bennett, G.N. Effect of ArcA and FNR on the expression of genes related to the oxygen regulation and the glycolysis pathway in Escherichia coli under microaerobic growth conditions. *Biotechnol Bioeng* **92**, 147-159 (2005).

58. Nam, H. Network context and selection in the evolution to enzyme specificity. *Science* **337**, 1101-1104 (2012).

59. Jenkins, L.S. & Nunn, W.D. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in Escherichia coli: the ato system. *Journal of Bacteriology* **169**, 42-52 (1987).

60. Magnuson, K., Jackowski, S., Rock, C.O. & Cronan, J.E. Regulation of fatty acid biosynthesis in Escherichia coli. *Microbiological Reviews* **57**, 522-542 (1993).

61. Dellomonaco, C., Clomburg, J.M., Miller, E.N. & Gonzalez, R. Engineered reversal of the beta-oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476**, 355-359 (2011).

62. Pizer, L.I. The Pathway and Control of Serine Biosynthesis in Escherichia coli. *Journal of Biological Chemistry* **238**, 3934-3944 (1963).

63. Georgellis, D., Kwon, O. & Lin, E.C. Quinones as the redox signal for the arc two-component system of bacteria. *Science* **292**, 2314-2316 (2001).

64. Sauer, U., Canonaco, F., Heri, S., Perrenoud, A. & Fischer, E. The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli. *J Biol Chem* **279**, 6613-6619 (2004).

65. Roe, A.J., O'Byrne, C., McLaggan, D. & Booth, I.R. Inhibition of Escherichia coli growth by acetic acid: a problem with methionine biosynthesis and homocysteine toxicity. *Microbiology* **148**, 2215-2222 (2002).

66. Tuite, N.L., Fraser, K.R. & O'Byrne C, P. Homocysteine toxicity in Escherichia coli is caused by a perturbation of branched-chain amino acid biosynthesis. *J Bacteriol* **187**, 4362-4371 (2005).

67. McLaggan, D., Naprstek, J., Buurman, E.T. & Epstein, W. Interdependence of K+ and glutamate accumulation during osmotic adaptation of Escherichia coli. *J Biol Chem* **269**, 1911-1917 (1994).

68. Herrick, J. & Sclavi, B. Ribonucleotide reductase and the regulation of DNA replication: an old story and an ancient heritage. *Mol Microbiol* **63**, 22-34 (2007).

69. Boston, T. & Atlung, T. FNR-Mediated Oxygen-Responsive Regulation of the nrdDG Operon of Escherichia coli. *Journal of Bacteriology* **185**, 5310-5313 (2003).

CHAPTER 4:

Fast Swinnex Filtration (FSF): A fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media

Abstract:

Liquid chromatography tandem mass spectrometry (LC-MS/MS) provides a powerful means to analyze intracellular metabolism. A prerequisite to accurate metabolomics analysis using LC-MS/MS is a robust sampling and extraction protocol. One unaddressed area in sampling is a detailed examination of a suitable method for anaerobic cultures grown in complex media. Given that a vast majority of bacteria are facultative or obligate anaerobes that grow to low biomass density and need to be cultured in complex media, a suitable sampling and extraction strategy for anaerobic cultures is needed. In this work, we develop a fast-filtration method using pressure-driven Swinnex® filters (FSF). We show that the method is fast enough to provide an accurate snapshot of intracellular metabolism, reduces matrix interference from the media to improve the number of compounds that can be detected, and is applicable to anaerobic and aerobic liquid cultures grown in a variety of culturing systems. Furthermore, we apply the fast filtration method to investigate differences in the absolute intracellular metabolite levels of anaerobic cultures grown in minimal and complex media.

Introduction:

Metabolomics has played an instrumental role in furthering our understanding of intracellular metabolism[1-10]. Liquid chromatography tandem mass spectrometry (LC-MS/MS) based methods provide a powerful approach to interrogate the metabolome by combining throughput and sensitivity [11-16]. A prerequisite to accurate metabolomics analysis using LC-MS/MS is the optimization of the sampling and extraction protocol [17, 18]. For intracellular metabolites with a turnover on the order of seconds or less, it must be fast enough to provide an accurate snapshot of metabolism, but also produce a suitable sample for analysis. For liquid cultures, meeting these demands is non-trivial, difficult to achieve, but critically important if meaningful data is to be generated.

Many bacteria are facultative or obligate anaerobes. Consequently, the ability to assay bacteria from anaerobic cultures to improve our understanding of their biochemistry for scientific, therapeutic, and industrial endeavors is highly relevant. Many of these bacteria require supplementation with complex nutrients such as yeast extract (YE), peptones, and blood components, among others, in order to be cultured in the lab. This presents unique challenges to LC-MS/MS-based metabolomics methods. Anaerobic cultures often reach a much lower biomass than aerobic cultures. For instance in our experience, stationary phase cultures of wild-type *E. coli* grown anaerobically in 4 g*L-1 of M9 minimal media reach a culture density of approximately 0.25 gDW*L-1 while stationary phase cultures of wild-type *E. coli* grown aerobically in the same media reach a culture

density of approximately 1.35 gDW*L-1. As the biomass of the culture decreases, the interference from media components increases. This is particularly problematic for anaerobic cultures where the amount of organic acids found in the culture medium that are produced by fermentation hamper the ability to accurately measure intracellular organic acid levels. The problem of media interference is exacerbated for cultures grown with supplementation. A common supplement for auxotrophic strains of *E. coli* is YE, which encompasses the water-soluble portion of autolyzed yeast. Often the amount of YE added to the growth medium can be an order of magnitude greater than the culture density itself. This makes accurate differentiation of intracellular from extracellular components inherently problematic if they are not fully removed.

Many strategies exist for removing and differentiating intracellular from extracellular components. These include fast filtration [5, 17, 19], fast centrifugation [15], and direct extraction either from liquid cultures, such as shake flasks [20] or pH controlled bioreactors [6, 21-23], or from cultures grown on filters [24, 25]. Previous studies have shown that the time required to perform fast filtration with a typical filtration setup and vacuum pump is sufficient for compounds that turnover less quickly, but is not sufficient for the physiologically important compounds that turnover in the time frame of seconds [17]. Fast centrifugation appears to quench metabolism in a timely manner, but its application when working in an anaerobic chamber or with anaerobic cultures does not appear viable. Direct extraction provides the fastest means to quench metabolism. Unfortunately, the organic solvents needed to quench metabolism cause the bacterial membrane to

become permeable, resulting in cell leakage and inaccurate measurement of the intracellular metabolome [17, 26, 27]. If the culture density is sufficient, such as in a pH controlled bioreactor, the sample can be directly extracted and dilutions can be employed in order to reduce matrix interference [19, 21]. By taking a measurement of the culture filtrate in parallel, the intracellular concentrations can be determined from the difference of the whole broth and filtrate [19, 21]. When the culture density is not sufficient to allow for dilutions, the direct extraction and application of the differential method can still be employed, but the number of compounds that can be analyzed accurately can be limited due to matrix interference [20]. As an alternative, samples can be grown directly on the filter used to extract the culture [18, 24]. However, this approach does not allow for multiple filtrate and/or broth samples to be taken from the same culture at different time-points or phases of growth. It also does not appear suitable for use with an anaerobic chamber (a popular culturing method) due to contamination of the chamber atmosphere with organic solvents if the extraction is performed in the chamber itself, or exposure to oxygen if the filter cultures are removed from the chamber prior to extraction.

Several automated devices have been constructed to assist in rapidly sampling liquid cultures from bioreactors [22, 28-31] as well as from flasks [20, 32]. While improving the reliability of rapid sampling, the devices are optimized for specific culture conditions, which limit their broad use. For researchers who culture cells under a wide array of culturing systems, a more flexible sampling system is needed. This is particularly true if samples need to be obtained from

cultures grown in an anaerobic environment or environmental samples need to be obtained from the field. Most devices are designed to be used in conjunction with a direct extraction of the liquid culture, which causes problems for low biomass cultures and cultures grown with supplementation. These problems include decreased column life-time, increased instrument maintenance, and reduced number of compounds that can be accurately quantified due to ion-suppression and media interference. Thus, an alternative sampling method that reduces the amount of media included with the cell biomass while still quenching metabolism fast enough (i.e., on an equivalent time-frame to that of direct extraction methods) to provide an accurate snap-shot of intracellular metabolite levels is needed.

In this work, we sought to develop a rapid sampling and extraction method that 1) can be applied to a wide range of liquid culturing systems and environments (including anaerobic environments), 2) that provides sufficient sampling and quenching speed to arrest cellular metabolism in order to provide an accurate snap-shop of the intracellular metabolome, and 3) that minimizes matrix-induced interference for accurate analysis by LC-MS/MS. We describe the steps taken to optimize a fast-filtration sampling and extraction method implemented using pressure-driven Swinnex® filters (FSF) to meet these goals and the resulting optimized method. We show that by working with a syringe filter, we are able to quench metabolism at a comparable rate to that of the direct extraction approach. Further, we show that the method is applicable to sampling liquid cultures from a variety of culturing vessels and conditions, and allows for

greater coverage of metabolites to be accurately quantified using LC-MS/MS. In addition, we apply the method to investigate differences in the absolute intracellular metabolite levels of anaerobic cultures grown in minimal media and media supplemented with YE.

Materials and Methods:

Chemicals and reagents:

Water, methanol, acetonitrile, acetonitrile + 0.1% formic acid, and water used for extraction were purchased from Honeywell Burdick & Jackson® (Muskegon, MI). Ammonium formate and triethylammonium acetate were purchased from Sigma-Aldrich (St. Louis, MO). Yeast extract was purchased from Fisher® Scientific (Pittsburgh, PA). Metabolically labeled internal standards were generated as described previously [20] from batch cultures of *E. coli* grown on uniformly labeled ^{13}C glucose, started from over-night pre-cultures of *E. coli* also grown on uniformly labeled ^{13}C glucose. Swinnex® filter holders and 0.45 uM filters (PES, mixed cellulose ester, and PVDF) were purchased from Millipore® (Billerica, MA).

Biological material and culture conditions:

E. coli K12 MG1655 (ATCC 700926), obtained from the American Type Culture Collection (Manassas, VA), were grown in 4 g/L glucose M9 minimal media [33] with trace elements [34] with or without 1 g/L of yeast extract. Growth under aerobic batch consisted of shake flasks in a water bath maintained at 37C and aerated at 500 RPM. Growth in an aerobic pH controlled bioreactors was used in both batch mode and in glucose limited continuous culture mode at two different dilutions rates (0.31 and 0.44 h-1) (see supplemental methods for more details of the chemostat experiments). The steady-state for glucose limited continuous cultures was achieved after 3-5 residence times and was verified by biomass measurements. Growth under anaerobic conditions consisted of shake

flasks in an anaerobic chamber (COY; 37 °C; 10% CO2, balance N2). Cultures were sampled during steady-state growth at an OD600 of 0.6 (aerobic batch), an OD600 of 1.0 (aerobic batch bioreactor), an OD600 of 3.0 (aerobic glucose limited chemostat) or at an OD600 of 0.3 (anaerobic batch). All batch culture samples were inoculated from overnight pre-cultures to a starting OD600 of approximately 0.01.

Sampling and extraction optimization:

2.0 mL of culture broth and culture filtrate were sampled and extracted using the FSF approach (Figure 1) or by direct injection into either pre-cooled organic solvent or liquid nitrogen when specified in the text. The extraction solvents used were 80:20 methanol:water pre-cooled to -80°C or 40:40:20 acetonitrile + 0.1% formic acid:methanol:water or 40:40:20 acetonitrile:methanol:water with or without buffer (as specified in the text) pre-cooled to -40°C. The volume of extraction solvent loaded into the syringe was 1.0 mL. For samples taken using the direct extraction approach, the extraction solvent was 4x that of the sample volume for whole broth and filtrate samples. Samples were then serially extracted twice with 200 uL of extraction solvent as described previously [20].

For anaerobic cultures, a filtrate sample for each replicate was used to calculate the external metabolite concentration. For aerobic cultures without supplementation, a pooled filtrate of replicates was found to be sufficient due to the larger fraction of biomass to media. In addition, it was found that after two rounds of directly extracting the filter and vortexing, subsequent rounds of

extraction of the biomass did not improve yields of metabolites. While true for *E. coli*, this would have to be reconfirmed for organisms with a different cellular membrane. The extracts were centrifuged at 16000 RPM at 4°C for 5 minutes. The supernatant was saved and the biomass was discarded. For acidic extraction solvents, the supernatant was neutralized with ammonium hydroxide (8 uL of 1 N ammonium hydroxide per 1 mL of extract containing 40% acetonitrile + 0.1% formic acid), centrifuged again at 16000 RPM at 4°C for 5 minutes, the supernatant saved and the precipitate discarded. Extracts were evaporated to dryness (Thermo Scientific™ Savant SpeedVac™, Waltham, MA), reconstituted in water, and stored in the -80°C until analysis. All extracts, extraction solvents, and filter disks contained in filter holders were kept on dry ice between vortexing, centrifugation, and pipetting steps.

LC-MS/MS analysis and quantification:

An XSELECT HSS XP 150 mm × 2.1 mm × 2.5 µm (Waters®, Milford, MA) with a UFLC XR HPLC (Shimadzu, Columbia, MD) was used for chromatographic separation. Mobile phase A was composed of 10 mM tributylamine (TBA), 10 mM acetic acid (pH 6.86), 5% methanol, and 2% 2-propanol; mobile phase B was 2-propanol. Oven temperature was 40°C. The chromatographic conditions are as follows: 0, 0, 0.4; 5, 0, 0.4; 9, 2, 0.4; 9.5, 6, 0.4; 11.5, 6, 0.4; 12, 11, 0.4; 13.5, 11, 0.4; 15.5, 28, 0.4; 16.5, 53, 0.15; 22.5, 53, 0.15; 23, 0, 0.15; 27, 0, 0.4; 33, 0, 0.4; (Total time [min], Eluent B [vol.%], Flow rate [mL*min-1])]. The autosampler temperature was 10°C and the injection volume was 10 uL with full loop injection. An AB SCIEX Qtrap® 5500 mass

spectrometer (AB SCIEX, Framingham, MA) was operated in negative mode. Electrospray ionization parameters were optimized for 0.4mL/min flow rate, and are as follows: electrospray voltage of -4500 V, temperature of 500 °C, curtain gas of 40, CAD gas of 12, and gas 1 and 2 of 50 and 50 psi, respectively. Analyzer parameters were optimized for each compound using manual tuning. The instrument was mass calibrated with a mixture of polypropylene glycol (PPG) standards.

Samples were acquired using the Analyst® 1.6.2 acquisition software and Scheduled MRM™ Algorithm (AB SCIEX). Integration was performed using MultiQuant™ 2.1.1 (AB SCIEX). IDMS [35, 36] with metabolically labeled internal standards was used for quantification. In brief, calibration curves of standards spiked with metabolically labeled internal standards were ran with each batch and used to back calculate the analyte levels in the whole broth and filtrate samples. The analyte levels in the samples were scaled to the amount of biomass in each culture determined at the time of sampling by optical density using the conversion factor of cell biomass to cell volume derived by Volkmer et al, 2011 [37] and experimentally derived conversion of cell density (gDW*L-1) to optical density of 0.45 for the used spectrophotometer. The differential method was then applied to the whole broth and filtrate samples [21] to derive the intracellular concentration. Linear regressions from calibration curves for compound quantification were based on peak height ratios and the logarithm of the concentration of calibrator concentrations from a minimum of four consecutive concentration ranges that showed minimal bias. A peak height

greater than 1e3 ion counts and signal to noise greater than 20 were used to define the lower limit of quantification (LLOQ). Quality controls and carry-over checks were included with each batch. Due to the number of biological isomers, the integration of each compound is manually checked.

Statistical analysis:

All statistical and correlation analyses were done using R [38] (R Development Core Team, 2011) or MetaboAnalyst [39].

Results and Discussion:

A fast filtration method based on pressure-driven Swinnex® filters (FSF) was explored for its suitability for use with anaerobic cultures, and compared to a direct extraction of the whole culture broth and culture filtrate and application of the differential method [21] (see materials and methods). An initial comparison between the direct extraction and application of the differential method and fast-filtration methods showed that the number of compounds that can be accurately assayed was increased when using FSF (Figure 2), but the levels of compounds that turn-over rapidly (e.g., ATP) were decreased, resulting in a low energy charge ratio (Supplemental Figure S1). Therefore, the first priority was to increase the speed at which metabolism was quenched using FSF.

Directly freezing the Swinnex® holder and filter in liquid nitrogen immediately after filtering the culture broth appeared to be a viable strategy to quickly quench metabolism. Unfortunately, cultures sampled using this approach were found to have a physiologically low energy charge ratio (Supplemental Figure S1). When directly extracting the culture broth and filtrate and applying the differential method using organic solvent was compared to freezing the broth and filtrate in liquid nitrogen and then extracting the frozen broth and filtrate with organic solvent, it was deduced that the intermediate step of freezing in liquid nitrogen was insufficient to quench metabolism on the time-scale required to obtain an accurate snap-shot of metabolism. From this finding, a means to expose the filtered biomass to the pre-cooled extraction solvent in a time-frame similar to that of the direct extraction approach was targeted. It was found that

by injecting organic solvent into the syringe housing and filter using a second syringe immediately after filtering the culture broth, a dramatic increase in the energy charge ratio could be obtained (Supplemental Figure S1).

The type of extraction solvent used and the addition of buffers to control for pH during the extraction process on compound stability were explored. In agreement with previous findings [18, 20], the combination of acetonitrile, methanol, and water was superior to the combination of only methanol and water for the more liable (i.e., phosphorylated) compounds (Supplemental Figure S2). The use of a buffered extraction solvent did not show any improvements in increasing the concentration of the more liable compounds (Supplemental Figure S3). Therefore, acidic acetonitrile was used as the extraction solvent for subsequent tests.

The material of the filter pad utilized in the extraction protocol was varied to understand its impact on the quality of the sample. The comparison of different filter materials revealed that NADH was increased in the cellulose and PVDF filters by 2.2 and 2.1 fold over the PES filters, respectively; NADPH was increased in the cellulose and PVDF filters by 1.2 and 1.4 fold over the PES filters, respectively. The loss of NADH and NADPH in the PES filters could have potentially been due to pi-pi bond interactions between the phenyl group of the PES and the niacin and/or adenine group of the NAD moiety. Also, the oxidized sulfone group could have interacted with the reduced niacin structure. The reduced levels of NADH and NADPH along with several other compounds (i.e., the nucleotide phosphates and glutathiones) with potentials for pi-pi bond

interactions resulted in a significant discrimination of the samples taken using the PES filters as determined by a partial least squares discriminatory analysis (PLS-DA) (Supplemental Figure S4). Based on these findings, and the fact that the cellulose filters displayed poor stability in organic solvent, the PVDF filters were used for subsequent tests.

FSF using various syringe sizes was compared to vacuum filtration to ensure that the use of pressure did not affect metabolism prior to extraction. Aside from physical theory, there is empirical evidence that shows that the pressure generated by lower volume syringes can be much greater than larger volume syringes (personal communication with Millipore®). Thus, samples taken using 5, 10, 20, and 60 mL syringes were compared to samples taken using vacuum filtration. The effect of syringe volume/pressure was found to be negligible. Indeed, similar levels of AMP, ADP, and ATP were found across all 5 conditions (Figure 3). Further, a detailed inspection of all metabolites assayed showed little difference between intracellular levels in each compound class across the different syringe sizes tested (Figure 4).

Measured differences between the FSF samples and samples taken using the direct extraction method and application of the differential approach for all compounds assayed showed little variation (Figure 4). Importantly, the energy charge ratio between the filtered and directly extracted samples were approximately equivalent, indicating the speed of quenching metabolism using the optimized fast-filtration was equivalent to that of the direct extraction method (Figure 3). The coverage of compounds was increased when using the FSF

method compared to the direct extraction method (Figure 4). The levels of more stable compounds (e.g., amino acids) are similar between the FSF method and direct extraction method (Figure 4). This indicates that the relative recovery of compounds (i.e., ratio of endogenous compounds to metabolically labeled internal standards) between the two methods is equivalent.

The absolute recovery of compounds using the FSF method was tested. The signal intensity of 98 compounds in a neat mixture without any manipulation, after a dry-down in a centrivap, after extraction using the direct extraction method, or after extraction using the FSF method were analyzed (Figure 5, Supplemental Figure S6). 24.5% of the compounds were found to have a significant different in signal intensity between either of the groups (n=8, P-value < 0.01, ANOVA). It was found that total signal intensity decreased from the neat mixture without any manipulation (1.18e7, 3.65e7), to the centrivap dry-down mixture (1.14e7, 3.57e7), to the direct extraction mixture (1.06e7, 3.46e7), and finally to the FSF mixture (9.54e6, 3.30e7) in both the significantly changed metabolites (Figure 5), and across all metabolites (Supplemental Figure SF6), respectively. The observed trend correlates with a decrease in signal intensity as the number of sample manipulation steps increased. A likely explanation for this trend is that as the number of sample manipulation steps is increased, a small amount of extracted material is lost. However, because the decrease in signal intensity does not exceed 20%, minimal affect on acquisition and no discernible effect on quantitation would be expected when using any of the extraction methods used in this study. In addition, the observed trend described above

would indicate that sample contamination or degradation is unlikely when using the extraction methods described in this study. Taken together, these results indicate that the final optimized filtration method was able to improve compound coverage and quench metabolism at a rate comparable to the direct extraction method for almost all compounds assayed. The optimized filtration method is shown in Figure 1.

The suitability of FSF to accurately, reliably, and quickly sample in aerobic or anaerobic conditions from batch or chemostat cultures was investigated by comparing relevant physiological ratios from wild-type *E. coli* samples grown in glucose minimal media (M9). The consistent ratios of energy charge (which would be expected for normal growing cells in glucose media regardless of the culture vessel or availability of oxygen) [40] were stable across the liquid cultures tested (Figure 3). The similar ratio of redox equivalents (i.e., NAD+, NADH, NADP+, and NADPH) within aerobic cultures and anaerobic cultures, but differing between aerobic and anaerobic cultures, provides additional evidence that the method can accurately, reliably, and quickly sample from a multitude of culture conditions (Figure 3). While not explored in this study, the method would be expected to allow for fast sampling from anaerobic bottle cultures, which are commonly used for strict anaerobic cultivation of microbes. The method would also be expected to fare well in the field, where environmental bacterial samples must be obtained.

Finally, in order to assess the impact of compound coverage on knowledge-gained per experiment, the optimized filtration method was applied to

investigate the metabolome of anaerobic cultures grown with and without YE supplementation. Whether using FSF or the direct extraction approach, analyses revealed that anaerobic cultures grown with YE supplementation have increased levels of intracellular amino acids and several mono-nucleotide phosphates most likely due to the uptake of amino acids and their precursors from the culture medium [41] (Figure 6). However, the number of amino acids and mono-nucleotide phosphates were higher in the samples taken by fast filtration, and the significant reduction in levels of the Glutathiones and Acetyl-CoA were masked for samples taken by direct extraction.

Twenty-two metabolites were quantifiable (see methods for cutoff) in both minimal and yeast extract samples using direct extraction, while 81 metabolites were quantifiable in both minimal and yeast extract samples using FSF. For the samples taken by the FSF, 15 compounds were found to be significantly changed between anaerobic growth with and without YE, 6 of which were amino acids and 4 of which were mono-nucleotide phosphates (Figure 6b). In contrast, for samples taken by the direct extraction approach, only 5 compounds were found to be significantly changed between the two conditions (Figure 6a). Three of the compounds, including 2 amino acids and NAD+, were common to both sampling approaches. The levels of glutamate and FAD were both found to be changed (increased and decreased, respectively) in the YE samples compared to the M9 samples for the samples taken by FSF, but they were not found to be significantly changed ($p<0.01$), nor did they have a fold change greater than two. The other amino acids, oxidized glutathione, ADP-glucose, UMP, dUMP, and 3'-

5’-cyclic GMP, could not be quantified in the YE samples for the samples taken by direct extraction. Acetyl-CoA and reduced glutathione were found to be elevated in the M9 samples compared to the YE samples for the samples taken by direct extraction, but their elevation was not significant ($p<0.01$). The fewer number of compounds that could be detected in the samples taken by direct extraction and the variances in the compounds between replicates resulted in a poorer discrimination between the two groups (Supplemental Figure S5) as determined by PLS-DA. This is exemplified by the change in axis scale between samples taken by fast filtration and direct extraction.

These results indicate that when sampling low biomass cultures supplemented with complex media, the two methods can provide overlapping findings, but the detail and breadth of those findings can be severely decreased if matrix reduction strategies are not employed. The reduced number of compounds and greater variance in the compounds measured when matrix reduction strategies were not employed limited downstream statistical and correlation analysis. The matrix interferences from the culture medium include salts and phosphate buffers that known to cause ion-suppression [19]. Besides directly lowering the detection limits via ion-suppression, increased on column matrix increases the base-line signal noise that interferes with the detection of low-abundant metabolites. The ability of the fast filtration method to provide a more suitable sample for analysis by LC-MS/MS compared to the direct extraction method allowed for more data points and information to be gained by the same experiment.

Conclusion:

LC-MS provides a powerful means to analyze intracellular metabolism. Unfortunately, the adverse effects of sample matrix can severely limit the information derived if matrix reduction steps are not included. These steps are a double edged sword in that matrix reduction steps often increase the time it takes to quench metabolism, which results in inaccurate metabolite levels for intracellular metabolites with fast turn-over times [19]. These difficulties are compounded further when working with anaerobic cultures, where metabolism must be quenched without the introduction of oxygen to the cells.

We have developed, validated, and described a fast-filtration method using pressure-driven Swinnex® filters (FSF) to overcome these challenges. The method provided fast sampling and quenching to obtain an accurate snapshot of metabolism. The method increased the coverage of compounds that can be detected by reducing matrix interference from the culture medium, which greatly improves the information that can be derived from a given metabolomics experiment. Because the method relies on pressure driven syringe filtration, it is flexible enough to sample anaerobic and aerobic liquid cultures grown in a variety of culturing systems. The developed method was applied to analyze and detail the metabolomes of *E. coli* when growing anaerobically in minimal and complex media containing yeast extract, and key differences were reported. It is envisioned that this sampling modality will provide researchers with a convenient means to obtain accurate intracellular (and simultaneously extracellular if the filtered medium is retained) metabolomics samples beyond those tested in this

study. Such samples could include environmental, anaerobic bottles, biofluids (e.g. blood and plasma), and samples from additional culturing conditions.

Acknowledgments

Chapter 4, in full, is a reformatted reprint of “Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media.” McCloskey, D, Utrilla, J, Naviaux, RK, Palsson BO, Feist AM. Metabolomics (2015) 11: 198. doi:10.1007/s11306-014-0686-2. The dissertation/thesis author was the primary investigator and author of the paper.

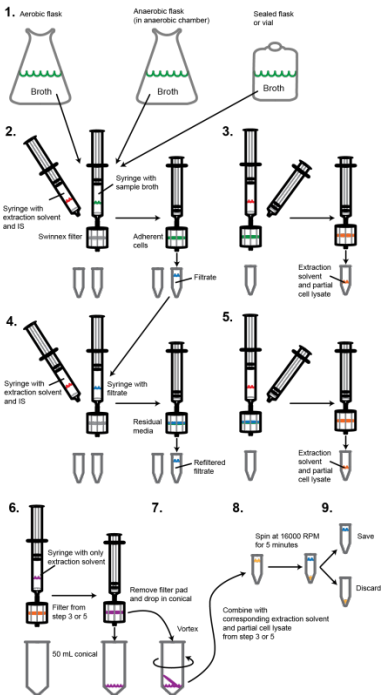

Figure 4.1: Fast filtration sampling and quenching using Swinnex® filters (FSF). 1) An accurate volume of culture broth was either sampled using a pipette and transferred to a syringe attached to a Swinnex® filter with the plunger removed (aerobic cultures) or collected using a syringe and 18.5 gauge blunt needle (anaerobic cultures). For the latter case, the plunger was then extended the volume of the syringe and then attached to a Swinnex® filter. Using a syringe volume that was a minimum of 2x greater than the liquid volume it was to contain allowed for a sufficient gas purge of the filter housing to remove residual culture or filtrate. In practice, we recommend using the largest syringe possible. 2) The cells were separated from the culture broth and retained on the Swinnex® filter pad by rapidly expelling the culture and extra volume gas through the filter housing and into a collection vessel. 3) The syringe was quickly removed, and a second syringe loaded with 1 mL of extraction solvent and labeled biomass pre-cooled to -40°C was quickly attached to the filter housing. The extraction solvent, labeled biomass, and extra volume gas was rapidly expelled through the filter into another collection vessel. The extraction solvent and partial cell lysate as well as the filter in the filter housing was stored in the -80°C for further extraction. The same procedure was repeated for each biological replicate. 4) The filtrate from step 2 for each replicate was filtered through a fresh Swinnex® filter, and 5) extracted as in step 3. The Swinnex® filter and extraction solvent were placed in the -80°C for further extraction. 6) The Swinnex® filter from step 3 or 5 was re-extracted with extraction solvent that does not contain internal standards. The eluent was collected in a 50 mL conical tube. 7) The filter holder was unscrewed over the 50 mL conical so that any residual extraction solvent would not be lost. The filter disk was removed and placed in the 50 mL conical using tweezers. The inside of the filter housing that is attached to the syringe was rinsed with a small volume of the extraction solvent from the 50 mL conical to remove any cells that were detached from the filter disk. The 50 mL conical with extraction solvent and filter disk were then vortexed for 30 seconds. 8) The extraction solvent and partial cell lysate from step 3 or 5 taken during the sampling procedure were added to the 50 mL conical and vortexed for an additional 30 seconds. The extraction solvent and cell lysate were then aliquoted into two eppindorf tubes, and the 50 mL conical and filter disk were discarded. 9) The cell debris was pelleted by spinning at 16000 RPM at 4°C for 5 minutes. The supernatant was saved in the -80°C for analysis and the cell debris was discarded. Further details of the FSF protocol are provided in the supplemental material.

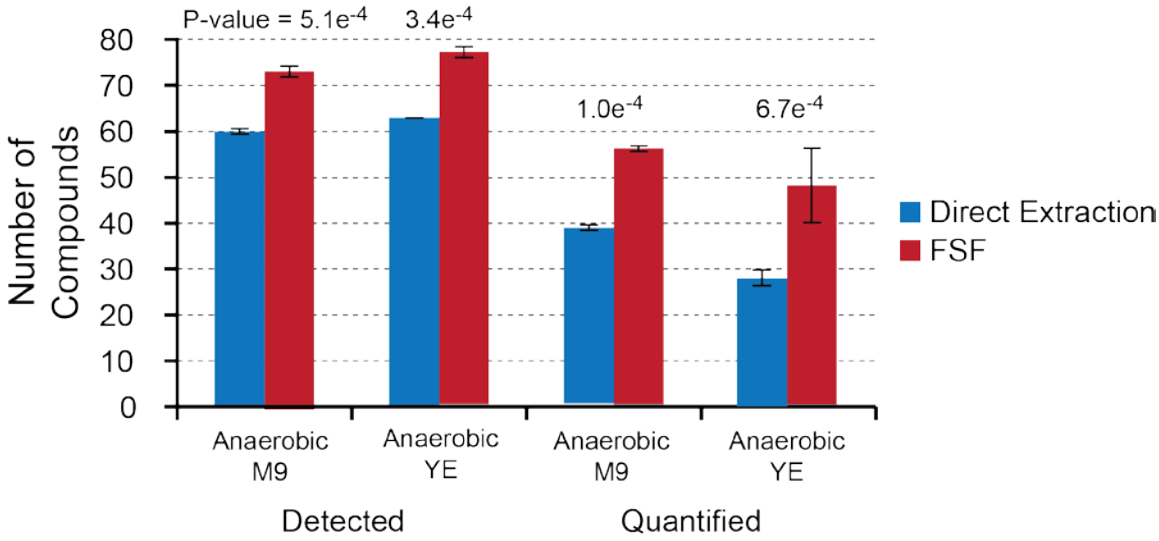

Figure 4.2: Sample matrix reduction by FSF. The number of compounds which could be detected with less than 50% signal contribution from the extracellular medium was higher in FSF samples compared to those obtained by direct extraction of the whole broth and application of the differential method. This was true for both anaerobic, wild-type *E. coli* grown on glucose M9 minimal media and anaerobic, wild-type *E. coli* grown on glucose M9 minimal media supplemented with 1 g*L-1 of YE. For the data shown, compounds that are considered 'quantifiable' are those that were found to have an average filtrate signal (n=3) of less than 50% compared to that found for the average intracellular filtration samples (n=3) or the average whole broth direct extraction samples (n=3) (i.e., $\frac{\textbf{filtrate}}{\textbf{whole broth}} * \mathbf{100}\% < \mathbf{50}\%$). Error bars represent standard deviations. The P-value (two-tailed Student's t-test) between the direct extraction and FSF are given above the bars.

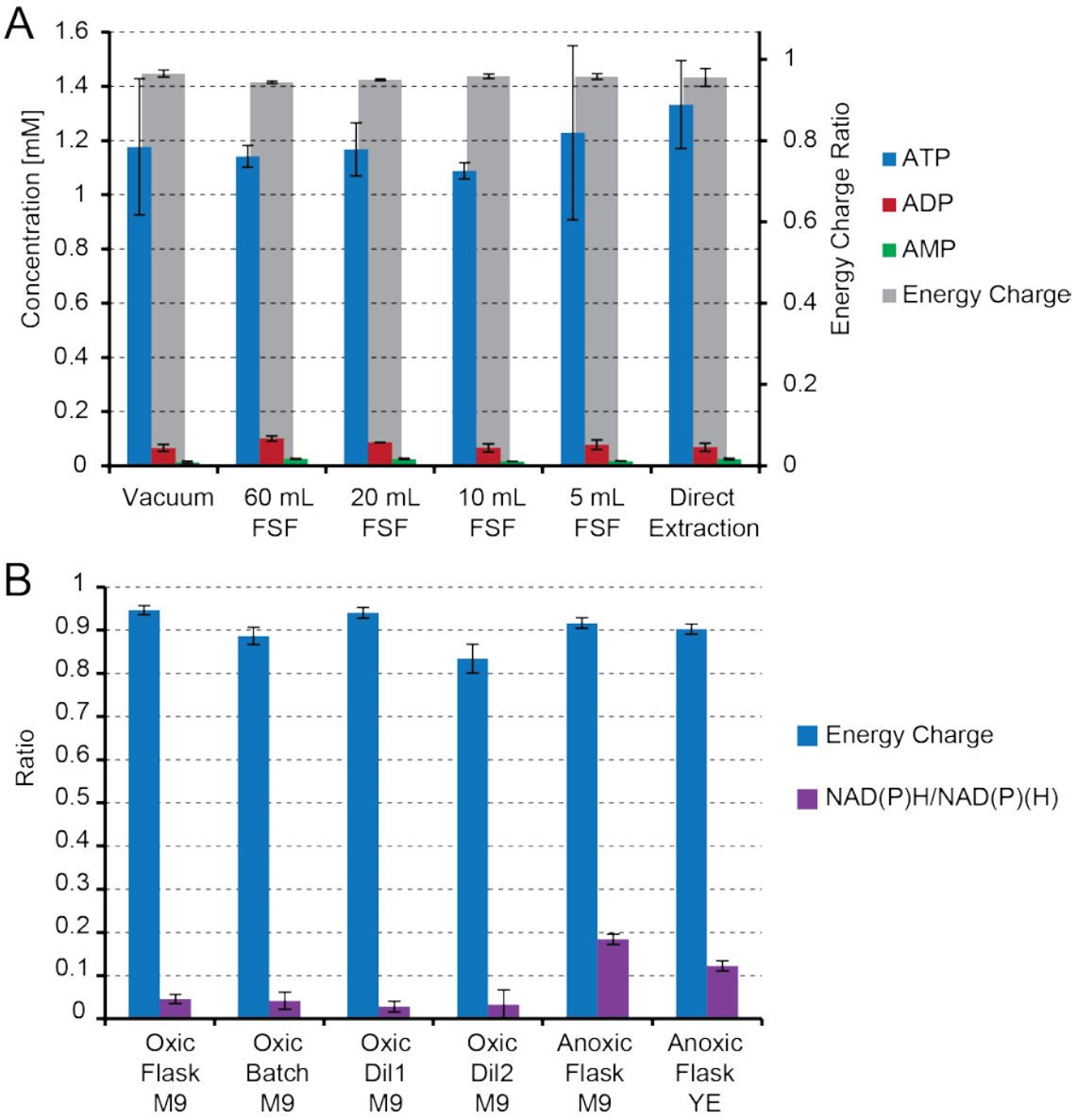

Figure 4.3: A) Comparison of intracellular ATP, ADP, AMP, and energy charge ratio (EC) between aerobic, wild-type *E. coli* grown in glucose minimal media. Cultures were sampled by vacuum filtration, by FSF using 5, 10, 20, and 60 mL syringes, and by direct extraction. Concentrations are averaged values (n≥3) in units of mM. B) Comparison of relevant intracellular physiological ratios for wild-type *E. coli* grown in glucose minimal media under aerobic batch growth (Oxic Flask M9), in a bioreactor during batch growth (Reactor Oxic Batch M9), in a bioreactor at two different dilution rates (Oxic Dil1 M9 and Oxic Dil2 M9), and wild-type *E. coli* grown in glucose minimal media under anaerobic batch growth without (Anoxic Flask M9) and with (Anoxic Flask YE) 1g*L-1 of yeast extract. Ratios were calculated from average concentration values (n≥3) in units of mM. Error bars represent standard deviations. Energy Charge = $\frac{\mathbf{ATP}+{}^{\boldsymbol{ADP}}/_{\mathbf{2}}}{\boldsymbol{ATP}+\boldsymbol{ADP}+\boldsymbol{AMP}}$ and NAD(P)H/NAD(P)(H) = $\frac{\mathbf{nadph}+\mathbf{nadh}}{\mathbf{nadph}+\mathbf{nadh}+\mathbf{nadp}+\mathbf{nad}}$.

	Vacuum	60 mL FSF	20 mL FSF	10 mL FSF	5 mL FSF	Direct Extraction
Amino Acids						
asn-L	2.72E-01	2.32E-01	1.70E-01	2.75E-01	2.97E-01	
citr-L	2.97E-02	3.51E-02	3.04E-02	3.42E-02	3.15E-02	
glu-L	9.82E+00	8.92E+00	8.15E+00	9.74E+00	9.79E+00	9.37E+00
gln-L	1.94E+00	1.81E+00	1.54E+00	1.98E+00	2.09E+00	2.46E+00
met-L	2.66E-02	2.22E-02	1.92E-02	2.19E-02	1.83E-02	2.00E-02
phe-L	1.77E-02	1.78E-02	1.33E-02	1.59E-02	1.52E-02	1.93E-02
ser-L	1.40E-01	1.62E-01	1.44E-01	1.51E-01	1.91E-01	
thr-L	2.31E-01	1.96E-01	1.98E-01	2.20E-01	2.38E-01	
trp-L	5.63E-03	5.33E-03	4.68E-03	5.57E-03	4.98E-03	5.39E-03
tyr-L	1.75E-02	1.50E-02	1.34E-02	1.67E-02	1.65E-02	2.91E-02
Cofactors and vitamins						
accoa	8.32E-01	8.38E-01	1.05E+00	8.90E-01	7.96E-01	1.19E+00
coa	9.39E-02	9.32E-02	8.23E-02	9.00E-02	9.30E-02	4.83E-02
fad	1.50E-02	1.39E-02	1.65E-02	1.36E-02	1.48E-02	1.36E-02
Nucleosides and bases						
ade	1.32E-03	3.78E-03	2.86E-03	1.70E-03	1.58E-03	
gua	4.42E-04			2.53E-04	3.03E-04	
gsn	3.46E-03	4.20E-03	3.99E-03	3.78E-03	3.68E-03	4.36E-04
hxan	8.16E-05	4.83E-05		3.18E-06		
ins	6.80E-04	9.74E-04	7.22E-04	7.00E-04	6.62E-04	
thym	2.75E-02	4.24E-02	3.58E-02	3.71E-02	3.67E-02	
Organic acids						
akg	5.36E-02	5.22E-02	5.47E-02	6.22E-02	5.60E-02	6.04E-02
5oxpro	1.01E-02	1.81E-02	1.53E-02	1.17E-02	1.28E-02	
acac	1.43E+00	1.54E+00	1.50E+00	1.56E+00	1.50E+00	
acon-C	1.17E-01	2.09E-01	1.73E-01	1.61E-01	1.70E-01	
glutacon	1.41E+00	1.84E+00	1.45E+00	1.53E+00	1.60E+00	
lac-L	2.75E-01			2.55E-01	2.18E-01	5.31E-01
mal-L	7.62E-01	8.17E-01	7.35E-01	7.83E-01	7.94E-01	8.40E-01
cit/icit	3.93E-01			1.90E-01	2.10E-01	
succ	2.53E-01	3.15E-01	2.63E-01	2.63E-01	2.65E-01	
icit	4.94E-03			3.60E-03	3.59E-03	
mmal	3.05E-02	3.84E-02	3.41E-02	3.22E-02	3.26E-02	
phpyr	6.16E-01	8.61E-01	7.44E-01	7.51E-01	7.88E-01	
NAD(P)(H)						
nadh	2.33E-03	7.51E-03	6.15E-03	9.21E-04	1.63E-03	9.03E-03
nadph	4.12E-02	3.54E-02	5.03E-02	4.45E-02	3.20E-02	1.02E-01
nad	2.77E-01	2.54E-01	2.24E-01	2.51E-01	2.44E-01	3.75E-01
nadp	8.51E-02	9.42E-02	7.52E-02	8.16E-02	9.17E-02	9.83E-02
gthrd	7.26E+00	7.73E+00	6.01E+00	7.62E+00	7.39E+00	8.00E+00
gthox	1.62E-01	1.56E-01	1.63E-01	1.59E-01	1.59E-01	3.46E-02

	Vacuum	60 mL FSF	20 mL FSF	10 mL FSF	5 mL FSF	Direct Extraction
Mono- and di-phosphorylated carbohydrates						
23dpg	1.57E-01	1.45E-01	1.29E-01	1.20E-01	9.90E-02	1.50E-01
fdp	8.76E-01	9.01E-01	7.87E-01	9.10E-01	7.05E-01	8.34E-01
6pgc	4.58E-02	1.92E-02	1.62E-02	2.81E-02	3.02E-02	
gam6p	3.84E-02	5.37E-02	5.22E-02	5.40E-02	5.56E-02	6.46E-02
g6p	1.14E+00	9.43E-01	8.35E-01	1.19E+00	1.03E+00	
glyc3p	1.72E-01	2.33E-01	2.08E-01	2.06E-01	2.16E-01	1.37E-01
pep	2.58E-02	3.51E-02	3.58E-02	3.69E-02	3.47E-02	3.25E-02
2pg/3pg	3.49E-01	4.26E-01	3.46E-01	4.29E-01	4.03E-01	1.95E-01
s7p	2.73E-02	2.30E-02	2.17E-02	2.87E-02	2.96E-02	
g1p	1.14E-01	1.48E-01	1.26E-01	1.38E-01	1.68E-01	
dhap	5.21E-01	5.69E-01	5.90E-01	5.97E-01	6.01E-01	3.05E-01
ru5p-D	9.94E-02	1.30E-01	1.23E-01	1.49E-01	1.36E-01	7.66E-02
NDPs						
adp	6.56E-02	1.01E-01	8.59E-02	6.62E-02	7.78E-02	6.82E-02
adpglc	3.30E-03			2.36E-03	2.66E-03	
dadp	5.58E-03	7.67E-03	7.92E-03	5.70E-03	6.52E-03	8.56E-03
dtdpglu	3.08E-03	9.01E-04	2.38E-03		1.99E-03	4.57E-03
gdp	4.51E-02	1.02E-01	1.03E-01	4.28E-02	5.89E-02	6.90E-02
udp	2.88E-02	4.81E-02	3.45E-02	2.38E-02	2.87E-02	3.09E-02
NMPs						
amp	1.33E-02	2.58E-02	2.55E-02	1.61E-02	1.76E-02	2.56E-02
camp	1.16E-02	1.97E-02	1.56E-02	1.60E-02	1.51E-02	
cmp	3.56E-03			2.14E-03	2.90E-03	
damp	2.25E-03	6.31E-03	5.93E-03	3.70E-03	3.58E-03	
dtmp	1.15E-02	1.34E-02	1.24E-02	1.23E-02	1.54E-02	4.57E-03
dump	2.22E-03	2.45E-03	2.57E-03	3.09E-03	3.29E-03	2.45E-03
gmp	7.53E-03	9.58E-03	1.11E-02	8.31E-03	8.03E-03	
imp	2.09E-02	2.10E-02	2.15E-02	2.41E-02	2.01E-02	4.32E-02
NTPs						
atp	1.24E+00	1.27E+00	1.20E+00	1.16E+00	1.35E+00	1.22E+00
ctp	1.52E+00	1.31E+00	1.09E+00	1.03E+00	1.43E+00	3.99E-01
datp	1.68E-01	1.74E-01	1.93E-01	1.82E-01	2.32E-01	
dctp	3.57E-02	2.83E-02	2.79E-02	3.23E-02	3.22E-02	4.61E-03
dttp	9.02E-02	9.05E-02	8.52E-02	8.16E-02	7.29E-02	1.94E-01
gtp	1.49E+00	1.59E+00	1.66E+00	1.49E+00	1.27E+00	
itp	1.34E-01	1.55E-01	1.47E-01	1.50E-01	1.43E-01	
utp	4.04E-01	4.19E-01	3.32E-01	3.46E-01	3.74E-01	3.92E-01

Figure 4.4: Heat map comparison of intracellular compounds grouped by compound class for aerobic, wild-type *E. coli* grown in glucose minimal media. Cultures were sampled by vacuum filtration, by FSF using 5, 10, 20, and 60 mL syringes, and by direct extraction and application of the differential method. We found that the metabolite levels for individual compounds are similar between the different approaches. In addition, it is evident that the cultures sampled using vacuum filtration or by FSF allow for quantification of more metabolites than when sampling using the direct extraction method. Metabolite abbreviations are given in Supplemental Table S1. Metabolite levels are based on averages (n≥3) in units of mM.

	ST	CE	DE	FSF	P-value	Fisher's LSD
glc-D	3.63E+04	4.98E+04	5.12E+04	1.65E+05	2.70E-06	FSF - CE; FSF - DE; FSF - ST
actp	4.92E+04	3.22E+04	3.33E+04	4.04E+04	6.55E-04	ST - CE; ST - DE
ade	1.36E+06	1.27E+06	1.23E+06	1.02E+06	5.21E-06	CE - FSF; DE - FSF; ST - FSF
akg	8.70E+04	1.07E+05	1.09E+05	1.98E+05	1.21E-08	FSF - CE; FSF - DE; FSF - ST
arg-L	1.41E+06	1.40E+06	1.33E+06	6.42E+05	3.69E-15	CE - FSF; DE - FSF; ST - FSF
asn-L	7.50E+05	7.24E+05	6.58E+05	4.99E+05	7.17E-08	CE - FSF; DE - FSF; ST - DE; ST - FSF
citr-L	4.52E+05	4.54E+05	4.43E+05	3.33E+05	2.01E-05	CE - FSF; DE - FSF; ST - FSF
coa	2.88E+05	2.21E+05	9.24E+04	7.58E+03	9.29E-10	CE - DE; CE - FSF; DE - FSF; ST - DE; ST - FSF
gln-L	8.67E+05	8.50E+05	7.69E+05	6.00E+05	8.00E-08	CE - FSF; DE - FSF; ST - DE; ST - FSF
glutacon	1.05E+05	8.97E+04	9.44E+04	7.85E+05	1.96E-13	FSF - CE; FSF - DE; FSF - ST
gthrd	2.85E+04	1.79E+04	8.60E+03	6.63E+02	5.37E-09	CE - DE; CE - FSF; ST - CE; ST - DE; ST - FSF
gua	1.96E+05	1.95E+05	1.87E+05	1.66E+05	3.86E-03	CE - FSF; ST - FSF
his-L	9.22E+05	9.18E+05	7.54E+05	5.54E+05	6.83E-08	CE - DE; CE - FSF; DE - FSF; ST - DE; ST - FSF
lac-L	9.36E+05	1.07E+06	1.27E+06	1.15E+06	2.98E-03	DE - ST
met-L	8.56E+05	8.26E+05	7.80E+05	6.08E+05	9.15E-06	CE - FSF; DE - FSF; ST - FSF
oaa	6.81E+05	4.21E+05	2.79E+05	3.19E+05	4.39E-08	CE - DE; ST - CE; ST - DE; ST - FSF
orn	2.35E+04	2.30E+04	2.11E+04	7.96E+03	1.45E-19	CE - FSF; DE - FSF; ST - DE; ST - FSF
oxa	2.90E+05	3.19E+05	3.42E+05	5.38E+05	4.63E-12	FSF - CE; FSF - DE; FSF - ST
phpyr	5.30E+04	4.36E+04	4.33E+04	4.70E+04	4.74E-03	ST - CE; ST - DE
pyr	3.03E+04	2.94E+04	2.89E+04	2.37E+04	3.52E-04	CE - FSF; DE - FSF; ST - FSF
ribflv	3.25E+05	2.76E+05	2.05E+05	2.21E+05	2.00E-04	CE - DE; ST - DE; ST - FSF
ser-L	4.12E+05	4.29E+05	4.08E+05	3.13E+05	2.24E-06	CE - FSF; DE - FSF; ST - FSF
succ	2.91E+05	3.22E+05	3.12E+05	4.47E+05	7.77E-05	FSF - CE; FSF - DE; FSF - ST
thr-L	1.32E+06	1.30E+06	1.13E+06	8.56E+05	3.92E-10	CE - DE; CE - FSF; DE - FSF; ST - DE; ST - FSF

Figure 4.5: Heat plot of the mean ion count (n=8) for significantly different metabolites (P-value < 0.01; ANOVA) in neat standard mixes that were extracted using different approaches. Neat standard mixes were analyzed without any manipulation (ST), analyzed after a dry-down in a centrivap and reconstituted in water (CE), analyzed after extraction using the direct extraction method (DE), or analyzed after extraction using the FSF method (FSF). The reconstitution volume for CE, DE, and FSF was the same as the initial volume of the neat standard mix. The mixes contained 98 representative intracellular metabolites, and were prepared at a concentration of moderate signal intensity for the instrument used. The full table of all 98 compounds is shown in supplemental figure S6. Extraction conditions that showed significant differences for a given metabolite (Fisher's Least Significant Difference (LSD)) are annotated next to the P-value.

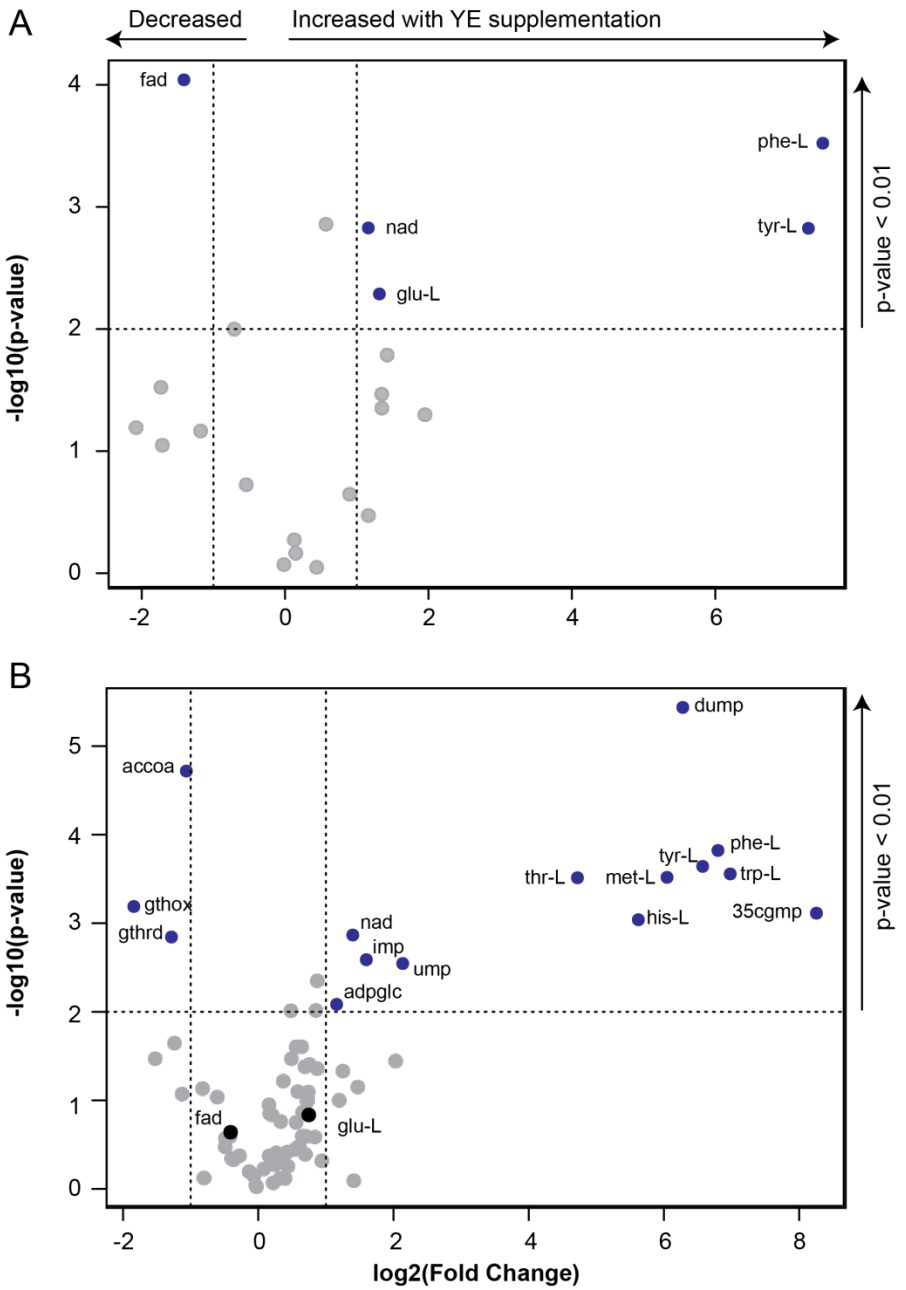

Figure 4.6: Volcano plot between wild-type anaerobic *E. coli* cultures grown in 4 g*L-1 M9 minimal media supplemented with or without 1 g*L-1 of yeast extract, and sampled by A) direct extraction or B) using the optimized FSF method. 22 metabolites were quantifiable (see methods for cutoff) in both minimal and yeast extract samples using direct extraction, while 81 metabolites were quantifiable in both minimal and yeast extract samples using FSF. Metabolites with a p-value greater than 0.01 and fold change greater than 2 are annotated on the plot (shown in blue). Out of the five metabolites that met the p-value cutoff of 0.01 in the direct extraction, three also met this criterion in the fast filtration measurement, and the remaining had p-values of 0.15 and 0.21 for glu-L and fad, respectively. The x-axis reflects the fold change between metabolites between the two conditions (i.e. log2(fold-change)). The y-axis reflects the significance (P-value; two-tailed Student's t-test) of the changes between the two conditions (i.e. -log10(p-value)).

References:

1. Nakahigashi, K. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol* **5**, 306 (2009).

2. Link, H., Kochanowski, K. & Sauer, U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat Biotechnol* **31**, 357-361 (2013).

3. Bennett, B.D. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat Chem Biol* **5**, 593-599 (2009).

4. Ibanez, A.J. Mass spectrometry-based metabolomics of single yeast cells. *Proc Natl Acad Sci U S A* **110**, 8790-8794 (2013).

5. Jozefczuk, S. Metabolomic and transcriptomic stress response of Escherichia coli. *Mol Syst Biol* **6**, 364 (2010).

6. Taymaz-Nikerel, H. Changes in substrate availability in Escherichia coli lead to rapid metabolite, flux and growth rate responses. *Metab Eng* **16**, 115-129 (2013).

7. Buescher, J.M. Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. *Science* **335**, 1099-1103 (2012).

8. Xu, Y.F., Amador-Noguez, D., Reaves, M.L., Feng, X.J. & Rabinowitz, J.D. Ultrasensitive regulation of anapleurosis via allosteric activation of PEP carboxylase. *Nat Chem Biol* **8**, 562-568 (2012).

9. Xu, Y.F. Regulation of yeast pyruvate kinase by ultrasensitive allostery independent of phosphorylation. *Mol Cell* **48**, 52-62 (2012).

10. Doucette, C.D., Schwab, D.J., Wingreen, N.S. & Rabinowitz, J.D. alpha-Ketoglutarate coordinates carbon and nitrogen utilization via enzyme I inhibition. *Nat Chem Biol* **7**, 894-901 (2011).

11. Bajad, S.U. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J Chromatogr A* **1125**, 76-88 (2006).

12. Cai, X. Analysis of highly polar metabolites in human plasma by ultra-performance hydrophilic interaction liquid chromatography coupled with

quadrupole-time of flight mass spectrometry. *Analytica Chimica Acta* **650**, 10-15 (2009).

13. van Dam, J.C. Analysis of glycolytic intermediates in Saccharomyces cerevisiae using anion exchange chromatography and electrospray ionization with tandem mass spectrometric detection. *Analytica Chimica Acta* **460**, 209-218 (2002).

14. Lu, W. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal Chem* **82**, 3212-3221 (2010).

15. Buescher, J.M., Moco, S., Sauer, U. & Zamboni, N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. *Anal Chem* **82**, 4403-4412 (2010).

16. Bennette, N.B., Eng, J.F. & Dismukes, G.C. An LC-MS-based chemical and analytical method for targeted metabolite quantification in the model cyanobacterium Synechococcus sp. PCC 7002. *Anal Chem* **83**, 3808-3816 (2011).

17. Bolten, C.J., Kiefer, P., Letisse, F., Portais, J.C. & Wittmann, C. Sampling for metabolome analysis of microorganisms. *Anal Chem* **79**, 3843-3849 (2007).

18. Kimball, E. & Rabinowitz, J.D. Identifying decomposition products in extracts of cellular metabolites. *Anal Biochem* **358**, 273-280 (2006).

19. Van Gulik, W.M. Fast sampling of the cellular metabolome. *Methods Mol Biol* **881**, 279-306 (2012).

20. McCloskey, D. A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. *Biotechnology and Bioengineering* **111**, 803-815 (2013).

21. Taymaz-Nikerel, H. Development and application of a differential method for reliable metabolome analysis in Escherichia coli. *Anal Biochem* **386**, 9-19 (2009).

22. De Mey, M. Catching prompt metabolite dynamics in Escherichia coli with the BioScope at oxygen rich conditions. *Metab Eng* **12**, 477-487 (2010).

23. Taymaz-Nikerel, H., van Gulik, W.M. & Heijnen, J.J. Escherichia coli responds with a rapid and large change in growth rate upon a shift from glucose-limited to glucose-excess conditions. *Metab Eng* **13**, 307-318 (2011).

24. Bennett, B.D., Yuan, J., Kimball, E.H. & Rabinowitz, J.D. Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. *Nat Protoc* **3**, 1299-1311 (2008).

25. Rabinowitz, J.D. & Kimball, E. Acidic acetonitrile for cellular metabolome extraction from Escherichia coli. *Anal Chem* **79**, 6167-6173 (2007).

26. Canelas, A. Leakage-free rapid quenching technique for yeast metabolomics. *Metabolomics* **4**, 226-239 (2008).

27. Link, H., Anselment, B. & Weuster-Botz, D. Leakage of adenylates during cold methanol/glycerol quenching of Escherichia coli. *Metabolomics* **4**, 240-247 (2008).

28. Lange, H.C. Improved rapid sampling for in vivo kinetics of intracellular metabolites in Saccharomyces cerevisiae. *Biotechnol Bioeng* **75**, 406-415 (2001).

29. Schaub, J., Schiesling, C., Reuss, M. & Dauner, M. Integrated sampling procedure for metabolome analysis. *Biotechnol Prog* **22**, 1434-1442 (2006).

30. Mashego, M.R., van Gulik, W.M., Vinke, J.L. & Heijnen, J.J. Critical evaluation of sampling techniques for residual glucose determination in carbon-limited chemostat culture of Saccharomyces cerevisiae. *Biotechnol Bioeng* **83**, 395-399 (2003).

31. Schaefer, U., Boos, W., Takors, R. & Weuster-Botz, D. Automated sampling device for monitoring intracellular metabolite dynamics. *Anal Biochem* **270**, 88-96 (1999).

32. Hiller, J., Franco-Lara, E., Papaioannou, V. & Weuster-Botz, D. Fast sampling and quenching procedures for microbial metabolic profiling. *Biotechnol Lett* **29**, 1161-1167 (2007).

33. Sambrook, J., and D. W. Russell Molecular cloning: a laboratory manual, 3rd ed., vol. A2.2. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001).

34. Fong, S.S. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

35. Mashego, M.R. MIRACLE: mass isotopomer ratio analysis of U-13C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. *Biotechnol Bioeng* **85**, 620-628 (2004).

36. Wu, L. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards. *Anal Biochem* **336**, 164-171 (2005).

37. Volkmer, B. & Heinemann, M. Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS ONE* **6**, e23126 (2011).

38. R Development Core Team (R Foundation for Statistical Computing, Vienna, Austria; 2011).

39. Xia, J., Psychogios, N., Young, N. & Wishart, D.S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* **37**, W652-660 (2009).

40. Cortassa, S. & Aon, M.A. Altered topoisomerase activities may be involved in the regulation of DNA supercoiling in aerobic-anaerobic transitions in Escherichia coli. *Mol Cell Biochem* **126**, 115-124 (1993).

41. Selvarasu, S. Characterizing Escherichia coli DH5alpha growth and metabolism in a complex medium using genome-scale flux analysis. *Biotechnol Bioeng* **102**, 923-934 (2009).

CHAPTER 5:

A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites

Abstract:

Comprehensive knowledge of intracellular biochemistry is needed to accurately understand, model, and manipulate metabolism for industrial and therapeutic applications. Quantitative metabolomics has been driven by advances in analytical instrumentation and can add valuable knowledge to the understanding of intracellular metabolism. Liquid chromatography coupled to mass spectrometry (LC-MS and LC-MS/MS) has become a reliable means with which to quantify a multitude of intracellular metabolites in parallel with great specificity and accuracy. This work details a method that builds and extends upon existing reverse phase ion-paring liquid chromatography methods for separation and detection of polar and anionic compounds that comprise key nodes of intracellular metabolism by optimizing pH and solvent composition. In addition, the presented method utilizes multiple scan types provided by hybrid instrumentation to improve confidence in compound identification. The developed method was validated for a broad coverage of polar and anionic metabolites of intracellular metabolism.

Introduction:

New manipulation tools and an improved understanding of intracellular metabolism have benefited industrial and medical bioengineering pursuits. The ability to model and manipulate the metabolism of microorganisms for the production of commodity chemicals from renewable resources has been demonstrated [1] and will continue to mature. The ability to better model and simulate human metabolism to develop new antimicrobials [2] and personalize patient treatment [3, 4] is an active area of health research.

A prerequisite of an accurate model is detailed knowledge of the underlying biochemistry. Quantitative metabolomics has contributed greatly to the body of knowledge of intracellular biochemistry. It involves the precise measurement of individual metabolite concentrations inside the cell, most often using isotope dilution mass spectrometry (IDMS) [5] (Fig.1). Since quantitative measurements are expressed in absolute amounts, this information can be integrated into biochemical models for a deeper investigation into metabolism. This approach has allowed for a comparison to Michaelis-Menten constants [6], in vivo calculations of reaction thermodynamics [7-10], construction and simulation of dynamic models of biochemical processes [11-15], novel discoveries of cellular function [16], enzyme activity [17], and allosteric regulation [18-22]. This information, when integrated with other omics data types, can be used to reveal different layers of cellular regulation [23-26].

Many advances in LC methods have improved the ability to separate and resolve the polar and charged species of intracellular metabolism. These

methods predominantly use hydrophilic-interaction-chromatography (HILIC) [27, 28], ion exchange (IE) [29], and reverse-phase ion-pairing (RIP) [30]. Some applications using aqueous normal phase (ANP) [31] and porous graphite carbon (PGC) [32-34] have emerged. Of the above-mentioned methods, ion-pairing methods have shown superior performance in terms of reproducibility, resolution, coverage, and sensitivity based on several head-to-head comparisons [35-37]. In particular, when compared with HILIC, ion-pairing methods have been found to have greater sensitivity due to improved separation of compounds that co-elute when using HILIC [35], and greater resolution of biological isomers (e.g., hexose phosphates) [38]. A notable drawback to the use of ion-pairing is its inability to analyze compounds in both positive and negative mode due to ion suppression [35, 39]. Consequently, it is often necessary to have a dedicated instrument in order to circumvent the need for lengthy cleaning cycles when switching to a method that uses the opposite polarity [35, 39]. Despite these caveats, ion-pairing methods using volatile ion-pairing agents and gradient elution [37-40] have enabled the separation of important biological isomers, provided a broad coverage of intracellular metabolites, and are MS compatible [35-37].

Advances in mass spectrometry instrumentation have improved the analyst's ability to detect, confirm, and accurately quantify the broad concentration ranges of compounds that comprise intracellular metabolism. Studies have shown that newer high resolution high mass accuracy mass spectrometers (HRMS), including time-of-flight (Q-TOF) and orbitrap type instruments, now offer sufficient linear range and sensitivity to fulfill the demands

of both qualitative and certain quantitative workflows [40, 41]. However, the QqQ remains the workhorse for targeted, quantitative analysis due to the greater linear range and sensitivity of single/multiple reaction monitoring scans (S/MRM) using triple quadrupole (QqQ or QTRAP) mass spectrometers [42, 43]. The large dynamic range of the QqQ is particularly important because the concentration ranges of intracellular metabolites can span >6 orders of magnitude.

In this work, a pH and solvent optimized RIP LC-MS/MS method for the quantification of intracellular metabolites was developed using an ABSCIEX QTRAP 5500 system, where the mass analyzer is a linear ion-trap (TRAP). First, a RIP-LC method is developed that builds on the extensive work of previous RIP methods to balance throughput and chromatographic resolution and improve sensitivity for complex intracellular matrices by modulating the solvent composition and pH of the mobile phases. Next, an acquisition method for targeted quantification is developed and validated. The quantification method leverages dual scan types to maximize coverage and improve confidence in compound detection. Finally, the stability and suitability for implementation in academic or industrial settings was demonstrated. The workflow utilizing the developed RIP LC-MS/MS for quantification of intracellular metabolites is shown in Figure 1.

Material and Methods:

Standards and reagents:

Standards were purchased from Sigma-Aldrich (St. Louis, MO) or Santa Cruz Biotechnology, Inc. (Dallas, TX). LC-MS reagents were purchased from Honeywell Burdick & Jackson® (Muskegon, MI). Metabolically labeled internal standards were generated as described previously [44] from batch cultures of *E. coli* grown on uniformly labeled ^{13}C glucose. Calibration standards were generated from stock solutions freshly prepared or kept in the -80C for no longer than 3 days. Calibration standards were then combined into mixes and aliquoted and lyophilized to dryness and stored at -80C. Aliquots were reconstituted in water, serially diluted, and spike with metabolically labeled internal standards to generate a calibration curve that spanned the lower and upper limits of detection for each compound.

Biological extracts:

E. coli:

Pooled or replicate samples of *E. coli* K12 MG1655 (ATCC 700926), obtained from the American Type Culture Collection (Manassas, VA), were grown in 4 g/L glucose or glycerol M9 minimal media [45] with trace elements [46] and sampled from a water bath that was maintained at 37 °C and aerated at 700 RPM. Samples were taken and extracted using a fast Swinnex® filtration approach described previously [47].

Red blood cells (RBCs):

Pooled red blood cell (RBC) samples were taken from freshly drawn blood of anonymous donors and extracted as described in the supplemental methods.

Instrumentation:

A XSELECT HSS XP 150 mm × 2.1 mm × 2.5 µm (Waters®, Milford, MA) with a Prominence UFLC XR HPLC (Shimadzu, Columbia, MD) was used for chromatographic separation. Mobile phase A was composed of 10 mM tributylamine (TBA), 10 mM acetic acid (pH 6.86), 5% methanol, and 2% 2-propanol; mobile phase B was 2-propanol. Oven temperature was 40°C. The chromatographic conditions for methods 1 and 2 are described in Table 1 and detailed in the supplemental methods. The autosampler temperature was 10°C and the injection volume was 10 uL with full loop injection. An AB SCIEX Qtrap® 5500 mass spectrometer (AB SCIEX, Framingham, MA) operated in negative mode with multiple reaction monitoring (MRM) was used for detection and quantification, with specific transitions shown in Table S-6. Where specified, the MRM was coupled to an information dependent acquisition (IDA) consisting of either an enhanced product ion (EPI) scan for confirmation of compound identity or an enhanced resolution (ER) and EPI scan for elucidation of compound isotopomer distribution. Electrospray ionization parameters were optimized for 0.4mL/min flow rate, and are as follows: electrospray voltage of -4500 V, temperature of 500 °C, curtain gas of 40, CAD gas of 12, and gas 1 and 2 of 50 and 50 psi, respectively. Analyzer parameters were optimized for each compound using manual tuning. The instrument was mass calibrated with a mixture of polypropylene glycol (PPG) standards.

Acquisition and quantification:

Samples were acquired using the scheduled MRM pro algorithm in Analyst® 1.6.2. Enhanced production ion (EPI) scans and enhanced resolution scans were extracted and processed using Analyst®. The compound library used for compound identification was generated by running sets of standards at a concentration range that allowed for good signal detection (i.e., greater than 1e5 cps). Samples were quantified using IDMS [5, 48] with metabolically labeled internal standards and processed using Multiquant® 2.1.1. Linear regressions for compound quantification were based on peak height ratios and the logarithm of the concentration of calibrator concentrations from a minimum of four consecutive concentration ranges that showed minimal bias. A peak height greater than 1e3 ion counts and signal to noise greater than 20 were used to define the lower limit of quantification (LLOQ). Quality controls and carry-over checks were included with each batch. Due to the number of biological isomers, the integration of each critical pair of compounds was manually checked.

Evaluation of RIP method performance:

Categories corresponding to a reduction in baseline noise, a reduction in isobaric interferences, a reduction in carryover for *E. coli* and RBC samples, improvement in isomer resolution, improvement in LLOQ, and linearity were used to evaluate the performance of the RIP methods presented in this study. A normalized score was calculated for each method for each category as described in the supplemental material to rank the performance of each method.

Results and Discussion:

Chromatographic optimization:

Prior to optimizing mass spectrometry acquisition parameters, a RIP method was sought that met the following criteria: (1) improved detection limits, and (2) balanced throughput and chromatographic resolution for intracellular metabolites. Two recently published RIP methods (each method is described in the supplemental methods section and in Tables 1 and 2), which are designated methods 1[40] and 2 [39], in addition to a third new method (method 3) described in Tables 1 and 2 were tested using the QTRAP LC-MS system. Method 3 builds upon methods 1 and 2, and incorporates additional optimizations to the mobile phases and gradient programs. Methods 1, 2, and 3 were compared with respect to their ability to meet the above criteria based on multiple chromatographic and detection parameters. The optimizations made to method 3 are first discussed, and then results of the comparison are presented. The results of the comparison are summarized in Table 3. Details of the calculations of the calculations of the comparison are expanded upon in Supplemental Tables S-2, S-3, S-4, and S-5, and in the supplemental material.

Mobile phase optimization:

Choice of solvent and composition was found to have a profound effect on lowering the baseline noise, reducing isobaric signals, decreasing carryover, and improving sensitivity of early eluting compounds. The baseline noise and isobaric interference for organic acids were found to be a function of the elution strength of mobile phase A (Figure S-1). By increasing the elution strength of

mobile phase A with the addition of 5% IPA, the baseline noise and isobaric signals were suppressed relative to weaker elution strength formulations of mobile phase A. An equivalent amount of methanol instead of IPA (empirically found to be 17%) was found to reproduce the results, indicating that it was not the choice of modifier or specific amount used, but most likely the elution strength of mobile phase A.

pH optimization:

In RIP chromatography, slight changes to the pH of mobile phase A caused by modulating the concentration of TBA or acetic acid can drastically alter the retention, selectivity, and signal intensity of most phosphorylated and carboxylated compounds. The work of Bennette et al, 2011 [37], explored the effects of pH on reverse phase ion-pairing chromatography, and found differences in signal intensity for phosphorylated compounds when modulating the acetic acid concentration by as little as 1 mM. We hypothesized that by increasing the pH closer to neutral, we could induce a second charge on the phosphate group (pKa ≈ 6). We reasoned that this would confer increased retention by attracting additional ion-pairing agents (Figures 2a and 2b). In addition, the decrease in acetic acid concentration[i] would confer increased retention by enhancing the absorption of ion-pairing agent to the stationary phase, which in turn would increase retention of oppositely charged solutes [49-51]. We also reasoned that the additional charge and decreased concentration of acetic acid might increase signal intensity by facilitating the negative ionization process through a pH change in the electrospray droplet [52-55]. During

evaporation of the electrospray droplet, the acetic acid would be removed faster than the TBA; this would increase the pH of the droplet, which could potentially further increase the percent charged state of the phosphorylated or carboxylated ion, resulting in greater signal intensity. Improved sensitivity for many phosphorylated and carboxylated compounds were found (Figure S-6 and Table S-3). It should be noted that improvements in peak shape [56] as well as improvements to the ionization process [57] from the use of IPA also appeared to play an important role in improving sensitivity as well.

For all phosphorylated and carboxylated compounds (with the exception of pyruvate, which decreased in retention), the retention time dramatically increased upon decreasing the concentration of acetic acid in mobile phase A from 15 mM (pH4.95) to 10 mM (pH6.86). This was noticeable for changes in as little as 0.5 mM in acetic acid concentration (Figures 2c and 2d). It was observed that a change in pH also changed the selectivity of glucose-, mannose-, and fructose-6-phosphate (g6p, man6p, and f6p) (Figure S-3). At a mobile phase pH of 4.9, fructose-6-phosphate eluted after galactose 1-phosphate (gal1p), and the retention time difference between g6p and man6p allowed for near baseline separation when using either the 169 or 199 trace. However, upon increasing the pH to 6.86, f6p eluted before gal1p, and the retention time difference between g6p and man6p decreased. Interestingly, the increased pH was found to separate additional pentose-phosphate isomers in a glucose-grown wild-type *E. coli* sample that were unresolved at a lower pH (Figure S-5). We have been

unable to purchase standards of sufficient purity to confirm the identity of the unknown pentose-phosphate isomers.

The change in selectivity between the hexose phosphate isomers and general improvements in sensitivity for phosphorylated and carboxylated compounds found when using a higher pH provides further insight into the mechanism by which the pH-induced differences were caused. We postulated that the increased pH above the pKa of the second acidic hydroxide moiety of the phosphate group allows for interaction with an additional ion-pairing agent. This would confer the increased retention that we have observed, but could also alter the solvation sphere that would interact with the polar end-capping or lessen its interaction, which would confer changes in selectivity between the isomers. While the interaction of pH, stationary phase, ion-pairing agent, counter-ion, and organic modifier on analyte selectivity, retention, and ionization efficiency is a complex and greatly debated topic [57-62], our findings provided additional insight into the mechanisms by which changes in pH, counter-ion concentration, and organic modifier can be modulated to improve selectivity and sensitivity in ion-pairing chromatography.

Comparison of RIP method performance:

Solvent blanks were injected using methods 1, 2, and 3 to assess baseline noise and determine if there were any isobaric signals present. A high baseline lowers the sensitivity of the assay by masking low abundant metabolites in background interference; and isobaric signals can confound the ability to measure a metabolite altogether. Methods 1 and 2 were observed to have high

baseline noise for most organic acids and were also observed to contain isobaric peaks corresponding to organic acid signals. In contrast, both the high baseline and isobaric signals corresponding to organic acids are greatly reduced in method 3.

Samples of RBC and *E. coli* extracts were ran at the same concentration and acquisition parameters followed by a series of blank injections using methods 1, 2, and 3 in order to assess carryover. Carryover can corrupt analysis and severely diminish daily sample throughput by requiring injections of additional blanks to not only monitor carryover, but completely eliminate carryover between samples. Method 1 was found to have carryover for compounds in RBC and *E. coli* samples, respectively, compared to none for methods 2 and 3 (Table 3, Supplemental Figure S-2). Extensive system decontamination revealed that inadequate washing of the column was to blame potentially due to weaker elution strength of mobile phase B. Method 1 employs 100% methanol in mobile phase B while methods 2 and 3 employ 100% 2-propanol. Extending the isocratic wash step of method 1 resulted in negligible improvements to carryover (data not shown). The difference in carryover found between using methanol and 2-propanol as the elution solvent indicates that for RIP methods aimed at separating compounds of intracellular metabolism run on modern instrumentation, traditional solvents such as methanol may be inadequate for complete displacement of the ion-pairing agent and negatively charged analytes from the column stationary phase. Consequently, a less polar solvent such as 2-propanol is needed.

Standards of representative biological isomers in a neat solution were in run in order to assess the balance between gradient run-time and resolution. The gradient-time and overall run-time of method 1 was shorter (15.5 and 25 min, respectively) compared to the gradient times and overall run-times of method 2 (19 and 36 min, respectively) and method 3 (16.5 and 33 min, respectively). However, methods 2 and 3 proved superior in their ability to resolve relevant biological isomers compared to method 1 (Table 3, Supplemental Figures S-3 and S-4). In particular, method 3 provided comparable separation of hexose and pentose-isomers that we were readily able to detect in RBC and *E. coli* extracts compared to method 2 and superior separation compared to method 1, while using an intermediate gradient time and overall run-time. This was achieved by modulating the mobile phase solvent composition and pH (as discussed previously). Improved separation of other biological isomers was also noted for method 3. AMP/dGMP, ADP/dGDP, and ATP/dGTP were found to be baseline resolved (Fig. 3b). The separation between the nucleotide mono-, di-, and triphosphates was not found when employing the previous methods. This separation is particularly beneficial in that isotope interference from a more abundant metabolite can mask the transition for a less abundant metabolite with a nominal mass difference of 1 amu less. Given the narrow mass distribution between many of the nucleotide mono-, di- and triphosphates and their discrepancy in biological concentration (e.g., ITP and ATP), it is not uncommon for us to see multiple peaks for a given ion current window.

Finally, calibrators composed of standards spiked with metabolically labeled *E. coli* biomass were run to determine the sensitivity and linearity of the different methods for performing quantitative analyses. All methods showed good linearity, but certain methods showed better sensitivity for certain classes of compounds (Table 3). A substantial improvement in sensitivity for many of the amino acids, nucleosides, and nucleotides that elute early in the gradient was found when using methods 2 and 3 as compared to method 1 (Table 3, Supplemental Figure S-6, and Supplemental Table S-1). This was most likely a result of the additional starting organic in mobile phase A, leading to a decrease in surface tension and improved ionization efficiency [56-58]. It was found that many phosphorylated and carboxylated compounds showed increased signal intensity and improved lower limits of quantification (LLOQ) using method 3 compared to methods 1 or 2 (Table 3 and Supplemental Table S-1). Overall, method 3 was found to have a lower limit of quantitation (LLOQ) that was superior in most cases to methods 1 and 2 for many of the amino acids, nucleosides, and nucleotides as well as phosphorylated and carboxylated compounds.

In summary, the results from the comparisons indicated that method 3 was best able to meet the criteria for sensitivity while balancing sample throughput and resolution when implemented on the QTRAP system (Table 1).

MS acquisition optimization:

In addition to optimizing throughput, resolution, and sensitivity on the chromatographic side, optimization for confidence in detection and coverage on

the mass spectrometer side was targeted by leveraging the multiple scan-types afforded by hybrid instrumentation. A minimum of a quantifying transition corresponding to the endogenous analyte and a transition corresponding to the heavy internal standard (most often a metabolically labeled heavy-carbon-13 analog) is needed for targeted quantitative workflows of intracellular metabolism. However, for more robust compound identification, a second qualifying transition corresponding to the endogenous analyte is used. The ratio of signal intensity of the quantifying transition to the qualifying transition is an intrinsic property of the molecule under the same ionization energy conditions and should remain constant between calibrators and samples, and can be used to confirm the identity of a compound. When a qualifying transition does not exist, or there is a major discrepancy in the signal intensity between the quantifying and qualifying transition, the use of signal ratios becomes problematic. To resolve this problem, a linear ion-trap was utilized to perform an additional confirmation using information dependent acquisition (IDA) enhanced product ion (EPI) scans (Fig. 4a). The EPI scan is a higher resolution full scan of the product ion spectra following fragmentation. The EPI acquisition is triggered when the signal intensity for a given quantifying and qualifying MRM reaches a pre-specified threshold. The product ion spectrum can then be matched against a spectral library to provide a second layer of compound identification (Fig. 3c). A spectral library was generated by injecting pure standards of each target compound and extracting the product ion scans.

It was found previously that calibrators ran using method 3 showed good linearity and LLOQ (Supplemental Table S-1). The reproducibility of the quantification acquisition method in two different matrices of different origin (i.e., red blood cells and *E. coli*) was tested. In general, for both sample matrices, the relative standard deviations for most compounds were less than 30% as determined from triplicate injections of each sample type (Table 4, Supplemental Table S-1). The coverage of metabolites that were detected in both sample matrices indicates that the utilized method has sufficient sensitivity and resolution to readily measure a broad spectrum of compounds that are of importance to intracellular metabolism in complex matrices (Fig. 3a and Table 4). The performance of calibrators acquired using the above quantification method, as well as the reproducibility measured using two complex sample matrices, demonstrated the ability of the method to enhance confidence in correct compound identification and provide accurate quantification of polar and anionic intracellular metabolites with a biomass of 10 µg or less on column.

LC-MS platform stability:

In order to test the stability of the LC-MS/MS system, the retention time during a system stress test for a neat standard solution was tracked. The system stress test consisted of varying the mobile phase A components +/- 5.0%, varying the tubing length +/- 6 inches, and several combinations of mobile phase A adjustments and tubing length. The largest %RSD found for any compound was less than 15.8%, which corresponds to a 95% confidence interval of 6.65 and 7.42 minutes (Fig. 5). Since this variability affects similar compounds

equally, the ability to resolve biological isomers such as ribose 5-phosphate and ribulose 5-phosphate (Fig. 4 inset) is not compromised. Retention time over the course of multiple batches of red blood cell and *E. coli* samples was also tracked. The retention time in relevant sample matrices was not found to be compromised (Fig. 5). The low retention time variability of a neat standard solution injected periodically throughout the course of the LC-MS/MS methods life-time demonstrates that the method is also stable through multiple batches of columns, guard columns, and samples, as well as routine system maintenance (Fig. 5, Figure S-7 and Tables S-7 and S-8).

Conclusion:

The ability to sensitively and accurately resolve many active biological isomers and charged species using LC-MS has arguably made it the premier platform to investigate intracellular metabolism. This fact is demonstrated by the plethora of high quality LC-MS-enabled studies that have emerged in recent years [14, 16, 18-21, 25]. This work describes the development and validation of a RIP-LC-MS/MS method for quantitative investigations of intracellular metabolism. The method was validated for a broad class of intracellular metabolites including nucleotides, nucleotide bases, amino acids, sugar phosphates, organic acids, nucleotide phosphates, nucleotide sugar phosphates, energy and redox metabolites (e.g., oxidized and reduced glutathione, NAD(P)(H), and FAD), acetyl-CoA, and several vitamins. The method builds and improves upon existing RIP-LC methods by optimizing solvent composition and solvent pH in order to improve the balance between method throughput and chromatographic resolution, while also improving sensitivity. A key feature of this method is the ability to leverage multiple scan-types provided by hybrid instruments (such as the QTRAP 5500) to improve the confidence with which compounds can be detected in complex sample matrices. The ability to perform multiple scan-types during a single run provides an efficient means to interrogate a sample more thoroughly and with greater confidence that can be extended to a variety of hybrid instruments. Finally, the method was also found to be stable and reliable for long-term use.

Acknowledgments

Chapter 5, in full, is a reformatted reprint of “A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites.” McCloskey, D, Gangoiti, JA, Palsson BO, Feist AM. Metabolomics (2015) 11: 1338. doi:10.1007/s11306-015-0790-y. The dissertation/thesis author was the primary investigator and author of the paper.

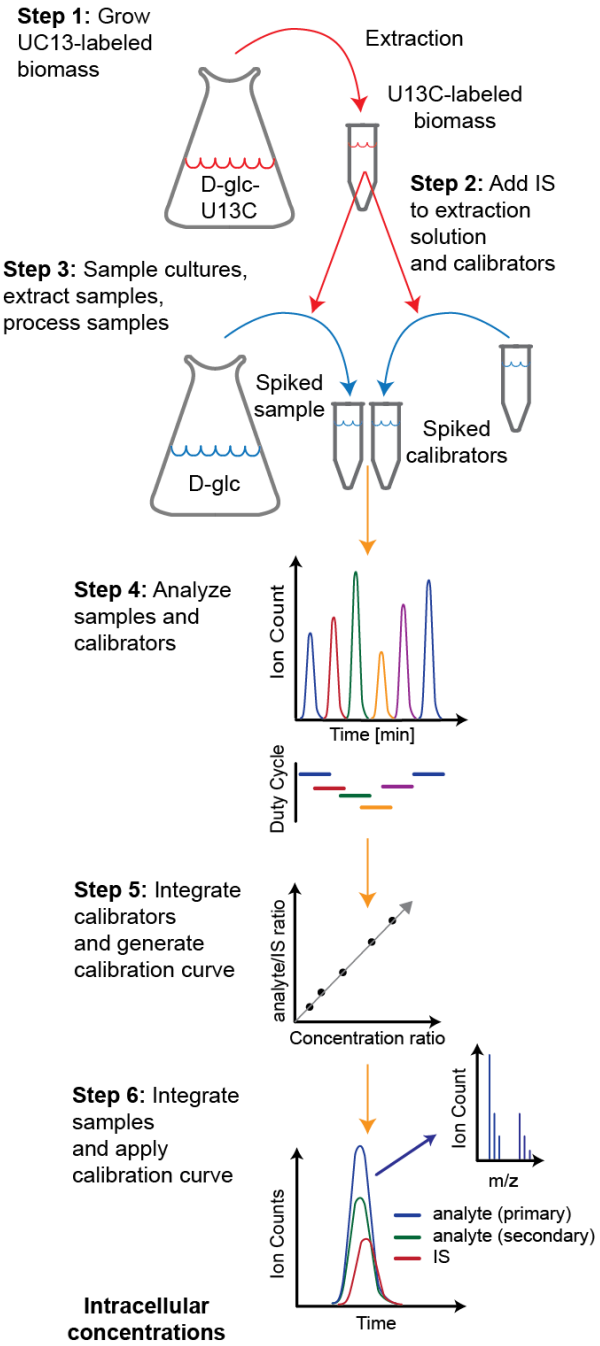

Figure 5.1: Overview of the LC-MS-enabled targeted absolute quantification workflow for investigations of intracellular metabolism using hybrid instrumentation. The absolute quantification workflow starts with rapidly sampling and extracting biological cultures with extraction solvent spiked with labeled standards. Labeled standards are also spiked into calibrators. Samples and calibrators are further processed and then metabolites are separated and acquired on the LC-MS instrument. A calibration curve is then constructed for each compound of interest in order to back-calculate the concentrations of analyte in the sample from the ratio of analyte to known amount of internal standard. Product ion spectra acquired for each of the analytes can be used to provide further confidence in compound identity.

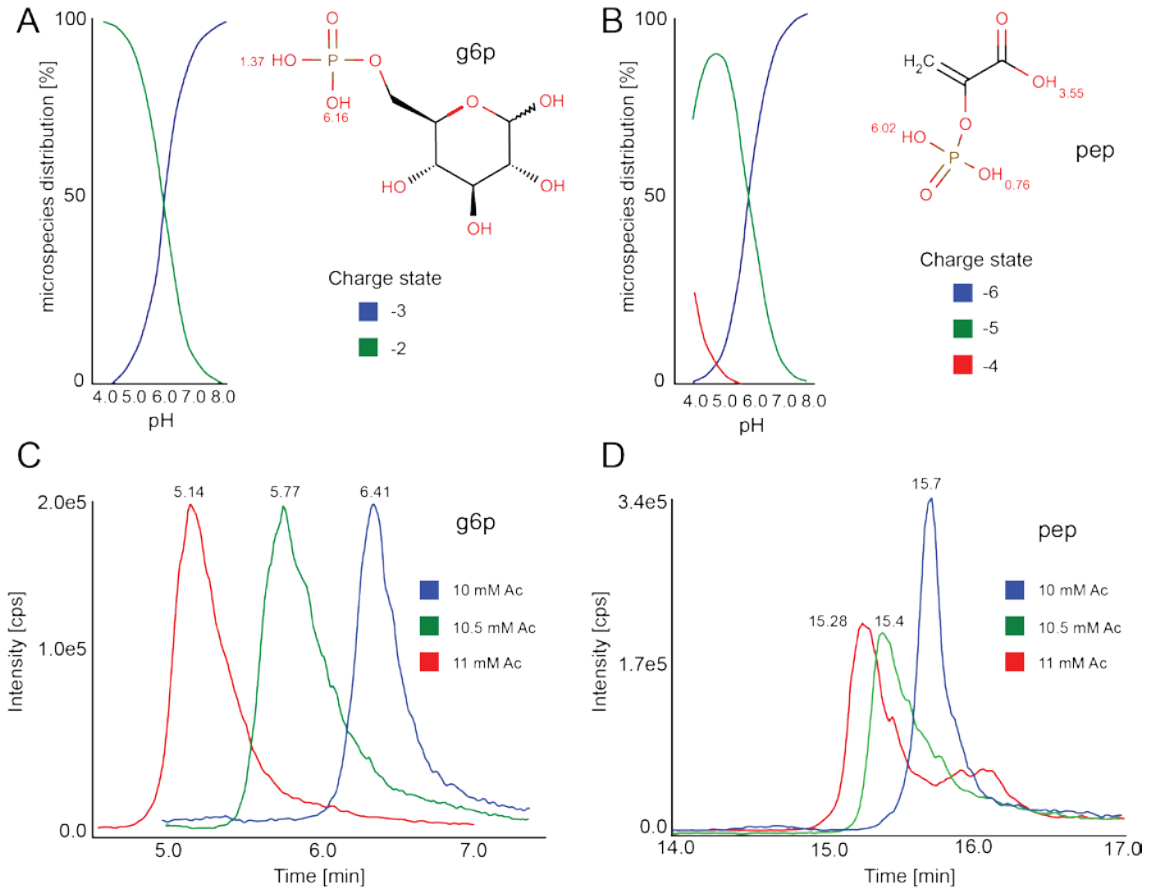

Figure 5.2: The effect of pH on ionization state and resulting change in retention time for glucose 6-phosphate (g6p) (A and C) and for phosphoenolpyruvate (pep) (B and D). We propose that by increasing the pH above the pKa of the second hydroxide moiety of the phosphate group, an increase in the dominant charge state of the compound in the solution as shown in the plot of microspecies distribution versus pH, and in the calculated pKa for the hydroxide moiety for g6p and pep (A and B) (see supplemental discussion). For the case of g6p, a dramatic change in retention time results (C); for the case of pep, a change in retention time and signal intensity results (D). A complete comparison of the resulting effect of signal intensity on the lower limits of quantification between the different methods is provided in supplemental table 1. The microspecies distribution and pKa were calculated using ChemAxon®.

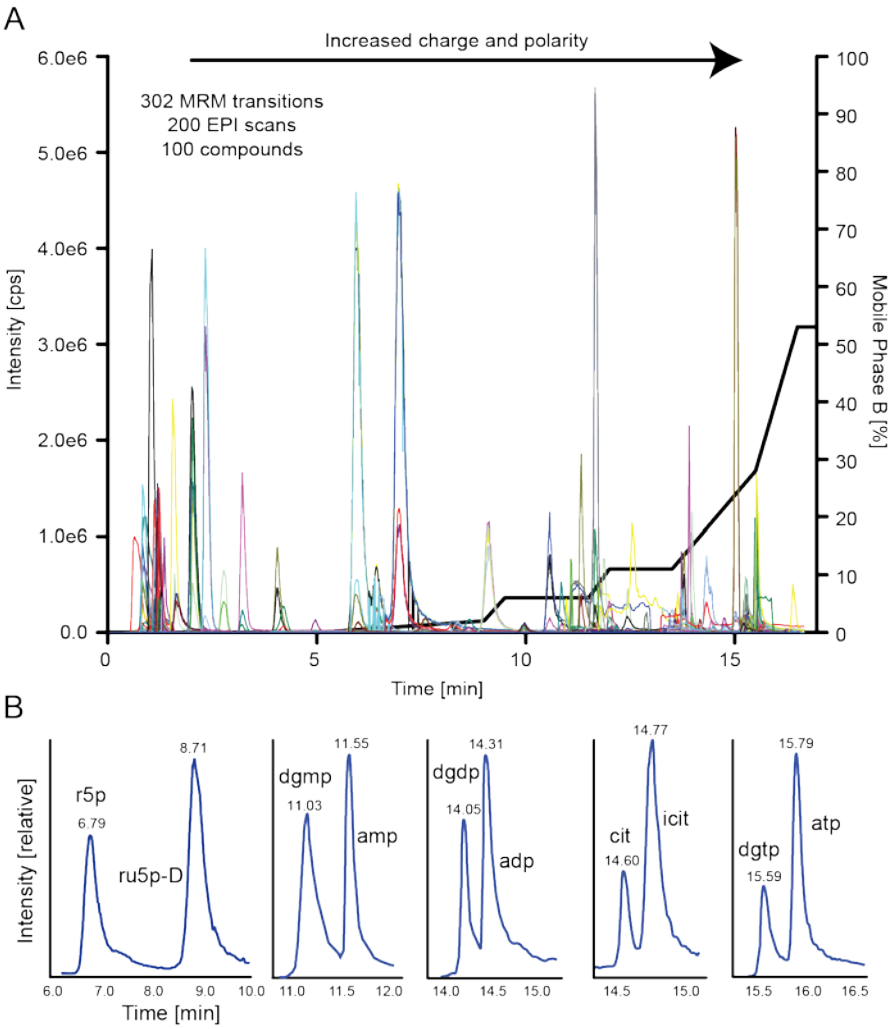

Figure 5.3: Example acquisitions using the LC-MS absolute quantification method. A) Representative injection of red blood cell extracts using the absolute quantitation acquisition method. The gradient profile is overlaid on the MRM traces. 302 transitions corresponding to 100 compounds are monitored in a single run. For most compounds, this includes the primary and secondary transition along with the uniformly labeled heavy carbon analog. B) Separation of biological isomers in calibrators. Shown are the baseline resolution of ribulose 5-phosphate (r5p) and ribulose-5-phosphate (ru5p-D), dGMP (dgmp) and AMP (amp), dGDP (dgdp) and ADP (adp), citrate (cit) and isocitrate (icit), and dGTP (dgtp) and ATP (atp).

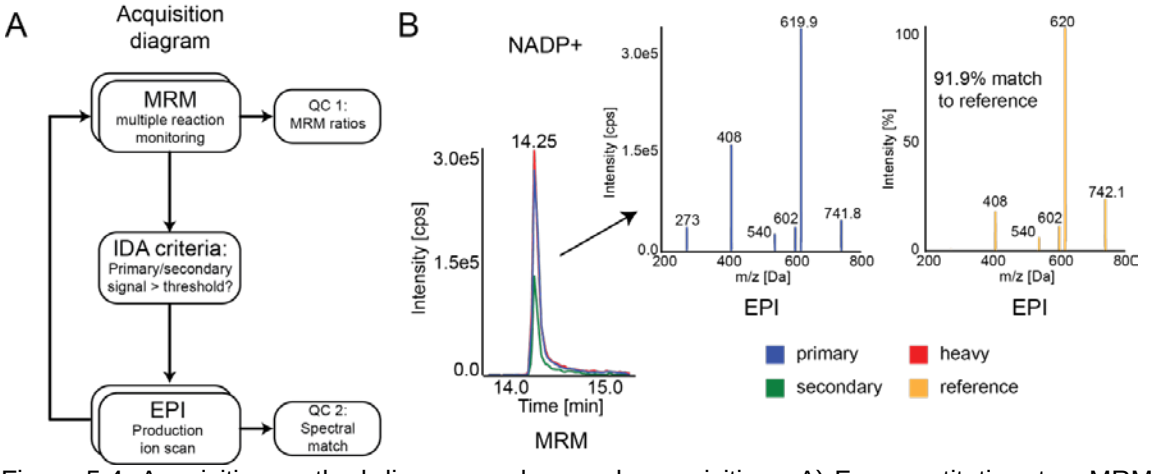

Figure 5.4: Acquisition method diagram and example acquisition. A) For quantitation, two MRM transitions (if two are available) are chosen per compound and scanned for during discrete retention windows during the LC gradient. When the primary and secondary transitions exceed a specified threshold, an information dependent acquisition (IDA) is triggered that utilizes an enhanced product ion (EPI) scan. The spectra of the EPI scan can then be compared to a library of product ion spectra for further compound identity confirmation. B) Acquisition of NADP+ (nadp) in an *E. coli* matrix using the quantitation method. The primary, secondary, and heavy transitions were acquired. The product ion scan taken from the primary transition is then compared to a reference library of product ion spectra. For this example, a 91.9% match of the endogenous analyte to the reference compound was found.

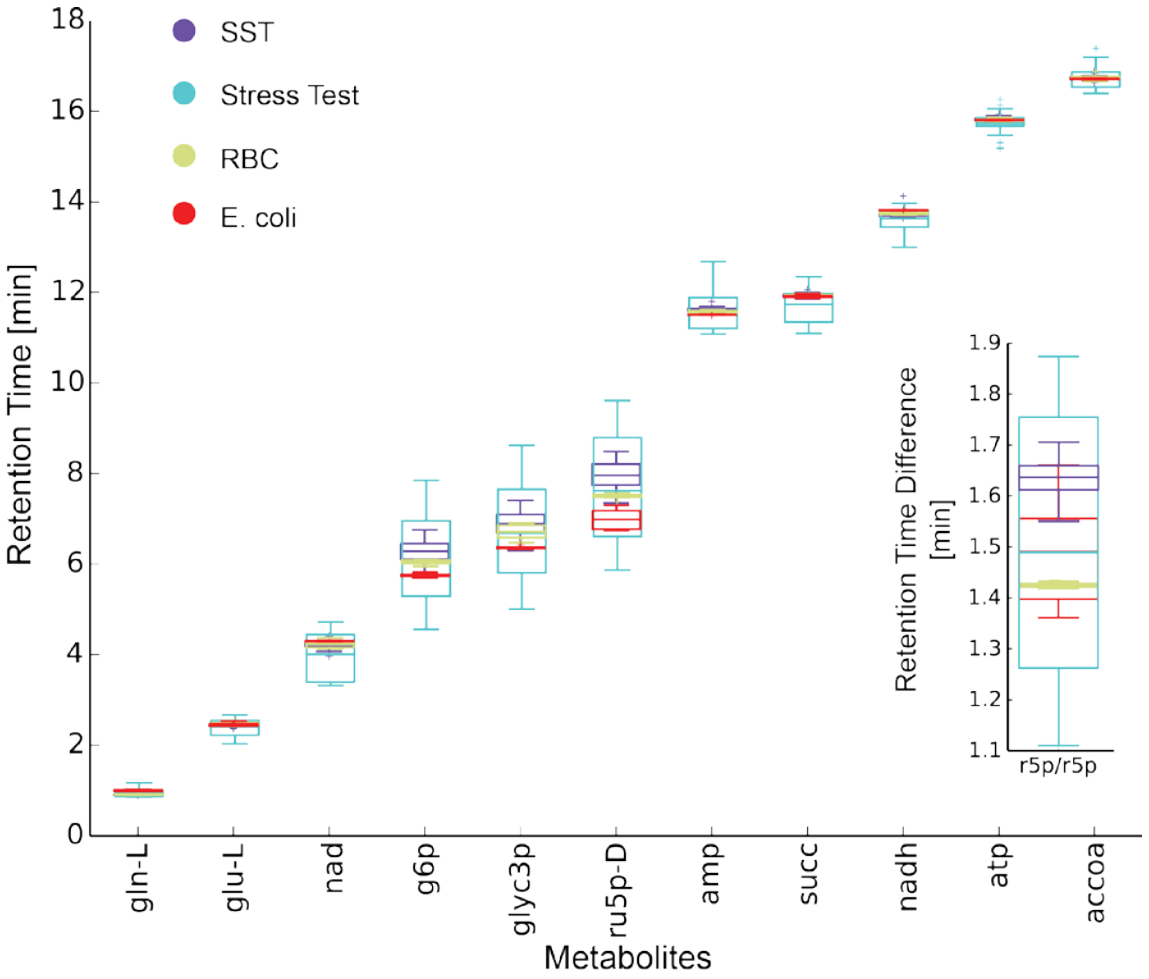

Figure 5.5: Retention time variability of representative compounds in a neat standard solution throughout the course of the methods life-time (SST), during a stress test (Stress Test), and in RBC (RBC) and *E. coli* (*E. coli*) samples. The box describes the mean and interquartile range. The whiskers describe the distribution of the data. The %RSD and 95% confidence intervals are given in Supplemental Tables S-6 and S-7. 4 columns from 3 different batches and 9 guard columns from 3 different batches have been cycled through during the life-time of the instrument. The stress test consisted of varying the mobile phase A components +/- 10.0%, varying the tubing length +/- 6 inches (approximately 10% of the internal LC tubing length), and several combinations of mobile phase A adjustments and tubing length. A pooled sample of RBCs was injected 6 times interspersed throughout a two week period of continuously running RBC samples. A pooled sample of *E. coli* was injected 6 times interspersed throughout a two week period of continuously running *E.coli* samples. The retention time differences of biological isomers in the inset are ribose 5-phosphate (r5p) and ribulose 5-phosphate (ru5p-D).

Table 5.1: Columns and chromatographic conditions used and compared in this study. Method 1 is based on the method described by Lu et al, 2010[40], method 2 is based on the method described by Buescher et al, 2010[39], and method 3 is described in this study. * Published method used a Waters® Atlantis T3 (150 mm × 2.1 mm × 1.8 µm)

Column	Dimensions and particle size	Mobile phase A	pH	Mobile phase B	Oven temperature	Method
Phenomenex® Synergi™ Hydro-RP	100 mm × 2.0 mm × 2.5 µm	10 mM TBA, 15 mM acetic acid, 3% methanol	4.95	Methanol	Ambient	1
Waters® XSELECT HSS XP*	150 mm × 2.1 mm × 2.5 µm	10 mM TBA, 15 mM acetic acid, 5% methanol	4.95	2-propanol	40 °C	2
Waters® XSELECT HSS XP	150 mm × 2.1 mm × 2.5 µm	10 mM TBA, 10 mM acetic acid, 5% methanol, 2% 2-propanol	6.86	2-propanol	40 °C	3

Table 5.2: Chromatographic gradients and flow rates used and compared in this study. Method 1 is based on the method described by Lu et al, 2010[40], method 2 is based on the method described by Buescher et al, 2010[39], and method 3 is described in this study.

Method 1		
Total time [min]	**Eluent B [vol.%]**	**Flow rate [mL*min-1]**
0	0	0.2
2.5	0	0.2
5	20	0.2
7.5	20	0.2
13	55	0.2
15.5	95	0.2
18.5	95	0.2
19	0	0.2
25	0	0.2

Method 2		
Total time [min]	**Eluent B [vol.%]**	**Flow rate [mL*min-1]**
0	0	0.4
5	0	0.4
10	2	0.4
11	9	0.35
16	9	0.25
18	25	0.25
19	50	0.15
25	50	0.15
26	0	0.15
32	0	0.4
36	0	0.4

Method 3		
Total time [min]	**Eluent B [vol.%]**	**Flow rate [mL*min-1]**
0	0	0.4
5	0	0.4
9	2	0.4
9.5	6	0.4
11.5	6	0.4
12	11	0.4
13.5	11	0.4
15.5	28	0.4
16.5	53	0.15
22.5	53	0.15
23	0	0.15
27	0	0.4
33	0	0.4

Table 5.3: RIP-LC Method comparison. Method 1 is based on the method described by Lu et al, 2010 [40], method 2 is based on the method described by Buescher et al, 2010 [39], and method 3 is described in this study. Explicit details of how the normalized score was derived for all criteria assessed are described in Supplemental Tables S-2, S-3, S-4, and S-5, and in the supplemental methods.

Criteria		**Normalized score**		
		Method 1	**Method 2**	**Method 3**
Blanks	Max baseline TIC	0.00	0.38	0.65
	Isobaric interferences	0.56	0.56	1.00
E. coli samples	Compounds w/o carryover	0.90	1.00	1.00
RBC samples	Compounds w/o carryover	0.91	1.00	1.00
Neat mixes	Resolution of isomers	0.19	0.63	0.88
	Run-time	1.00	0.56	0.68
Calibrators (LLOQ) improvement for amino acids, nucleosides, and nucleotides	*vs. Method 1*		*0.64*	*0.89*
	vs. Method 2	*0.36*		*0.79*
	vs. Method 3	*0.11*	*0.21*	
Sub-total:	Normalized sub-score	0.24	0.43	0.84
Calibrators (LLOQ) improvement for phosphate containing compounds	*vs. Method 1*		*0.71*	*0.83*
	vs. Method 2	*0.29*		*0.58*
	vs. Method 3	*0.17*	*0.42*	
Sub-total:	Normalized sub-score	0.23	0.57	0.71
Calibrators (LLOQ) improvement for organic acids	*vs. Method 1*		*0.37*	*0.67*
	vs. Method 2	*0.63*		*0.75*
	vs. Method 3	*0.33*	*0.25*	
Sub-total:	Normalized sub-score	0.48	0.31	0.71
Calibrators (Linearity)	Compounds w/ $R^2 > 0.98$	0.98	0.93	0.98
Total:	**Normalized total score**	**0.55**	**0.64**	**0.84**

Table 5.4: The total number of intracellular compounds that can be quantified and the total number of intracellular metabolites quantified with a relative standard deviation (RSD) less than 30% (n=3) in two representative matrices.

Red Blood Cells*		***E. coli*^**	
Quantified	RSD < 30%	Quantified	RSD < 30%
81	67	77	72

* Red blood cells were extracted and processed as described in the material and methods. ^ *E. coli* cells were grown on glycerol minimal media and extracted and processed as described in the material and methods.

References:

1. Yim, H. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* **7**, 445-452 (2011).

2. Shen, Y. Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc Natl Acad Sci U S A* **107**, 1082-1087 (2010).

3. Bordbar, A., Jamshidi, N. & Palsson, B. iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. *BMC Systems Biology* **5**, 110 (2011).

4. Gille, C. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* **6**, 411 (2010).

5. Wu, L. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards. *Anal Biochem* **336**, 164-171 (2005).

6. Bennett, B.D. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat Chem Biol* **5**, 593-599 (2009).

7. Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys J* **92**, 1792-1805 (2007).

8. Zamboni, N., Kummel, A. & Heinemann, M. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9**, 199 (2008).

9. McCloskey, D. A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. *Biotechnology and Bioengineering*, n/a-n/a (2013).

10. Noor, E., Haraldsdottir, H.S., Milo, R. & Fleming, R.M. Consistent estimation of Gibbs energy using component contributions. *PLoS Comput Biol* **9**, e1003098 (2013).

11. Taymaz-Nikerel, H., Borujeni, A.E., Verheijen, P.J., Heijnen, J.J. & van Gulik, W.M. Genome-derived minimal metabolic models for Escherichia coli MG1655 with estimated in vivo respiratory ATP stoichiometry. *Biotechnol Bioeng* **107**, 369-381 (2010).

12. Taymaz-Nikerel, H. Changes in substrate availability in Escherichia coli lead to rapid metabolite, flux and growth rate responses. *Metab Eng* **16**, 115-129 (2013).

13. Taymaz-Nikerel, H., van Gulik, W.M. & Heijnen, J.J. Escherichia coli responds with a rapid and large change in growth rate upon a shift from glucose-limited to glucose-excess conditions. *Metab Eng* **13**, 307-318 (2011).

14. Doucette, C.D., Schwab, D.J., Wingreen, N.S. & Rabinowitz, J.D. alpha-Ketoglutarate coordinates carbon and nitrogen utilization via enzyme I inhibition. *Nat Chem Biol* **7**, 894-901 (2011).

15. Chassagnole, C., Noisommit-Rizzi, N., Schmid, J.W., Mauch, K. & Reuss, M. Dynamic modeling of the central carbon metabolism of Escherichia coli. *Biotechnol Bioeng* **79**, 53-73 (2002).

16. Tepper, N. Steady-State Metabolite Concentrations Reflect a Balance between Maximizing Enzyme Efficiency and Minimizing Total Metabolite Load. *PLoS ONE* **8**, e75370 (2013).

17. Nakahigashi, K. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol* **5**, 306 (2009).

18. Link, H., Kochanowski, K. & Sauer, U. Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nat Biotechnol* **31**, 357-361 (2013).

19. Xu, Y.F., Amador-Noguez, D., Reaves, M.L., Feng, X.J. & Rabinowitz, J.D. Ultrasensitive regulation of anapleurosis via allosteric activation of PEP carboxylase. *Nat Chem Biol* **8**, 562-568 (2012).

20. Xu, Y.F. Regulation of yeast pyruvate kinase by ultrasensitive allostery independent of phosphorylation. *Mol Cell* **48**, 52-62 (2012).

21. Kochanowski, K. Functioning of a metabolic flux sensor in Escherichia coli. *Proc Natl Acad Sci U S A* **110**, 1130-1135 (2013).

22. Yuan, J. Metabolomics-driven quantitative analysis of ammonia assimilation in E. coli. *Mol Syst Biol* **5**, 302 (2009).

23. Cakir, T. Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol Syst Biol* **2**, 50 (2006).

24. Toya, Y., Nakahigashi, K., Tomita, M. & Shimizu, K. Metabolic regulation analysis of wild-type and arcA mutant Escherichia coli under nitrate conditions using different levels of omics data. *Mol Biosyst* **8**, 2593-2604 (2012).

25. Buescher, J.M. Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. *Science* **335**, 1099-1103 (2012).

26. Fendt, S.M. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol Syst Biol* **6**, 356 (2010).

27. Bajad, S.U. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J Chromatogr A* **1125**, 76-88 (2006).

28. Cai, X. Analysis of highly polar metabolites in human plasma by ultra-performance hydrophilic interaction liquid chromatography coupled with quadrupole-time of flight mass spectrometry. *Analytica Chimica Acta* **650**, 10-15 (2009).

29. van Dam, J.C. Analysis of glycolytic intermediates in Saccharomyces cerevisiae using anion exchange chromatography and electrospray ionization with tandem mass spectrometric detection. *Analytica Chimica Acta* **460**, 209-218 (2002).

30. Coulier, L. Simultaneous quantitative analysis of metabolites using ion-pair liquid chromatography-electrospray ionization mass spectrometry. *Anal Chem* **78**, 6573-6582 (2006).

31. Matyska, M.T., Pesek, J.J., Duley, J., Zamzami, M. & Fischer, S.M. Aqueous normal phase retention of nucleotides on silica hydride-based columns: method development strategies for analytes relevant in clinical analysis. *J Sep Sci* **33**, 930-938 (2010).

32. Antonio, C. Quantification of sugars and sugar phosphates in Arabidopsis thaliana tissues using porous graphitic carbon liquid chromatography-electrospray ionization mass spectrometry. *J Chromatogr A* **1172**, 170-178 (2007).

33. Pabst, M. Nucleotide and nucleotide sugar analysis by liquid chromatography-electrospray ionization-mass spectrometry on surface-conditioned porous graphitic carbon. *Anal Chem* **82**, 9782-9788 (2010).

34. Xing, J., Apedo, A., Tymiak, A. & Zhao, N. Liquid chromatographic analysis of nucleosides and their mono-, di- and triphosphates using porous graphitic carbon stationary phase coupled with electrospray mass spectrometry. *Rapid Commun Mass Spectrom* **18**, 1599-1606 (2004).

35. Lu, W., Bennett, B.D. & Rabinowitz, J.D. Analytical strategies for LC-MS-based targeted metabolomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **871**, 236-242 (2008).

36. Buscher, J.M., Czernik, D., Ewald, J.C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal Chem* **81**, 2135-2143 (2009).

37. Bennette, N.B., Eng, J.F. & Dismukes, G.C. An LC-MS-based chemical and analytical method for targeted metabolite quantification in the model cyanobacterium Synechococcus sp. PCC 7002. *Anal Chem* **83**, 3808-3816 (2011).

38. Luo, B., Groenke, K., Takors, R., Wandrey, C. & Oldiges, M. Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography-mass spectrometry. *J Chromatogr A* **1147**, 153-164 (2007).

39. Buescher, J.M., Moco, S., Sauer, U. & Zamboni, N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. *Anal Chem* **82**, 4403-4412 (2010).

40. Lu, W. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal Chem* **82**, 3212-3221 (2010).

41. Gertsman, I., Gangoiti, J.A. & Barshop, B.A. Validation of a dual LC-HRMS platform for clinical metabolic diagnosis in serum, bridging quantitative analysis and untargeted metabolomics. *Metabolomics* **10**, 312-323 (2014).

42. Pozo, O.J. Comparison between triple quadrupole, time of flight and hybrid quadrupole time of flight analysers coupled to liquid chromatography for the detection of anabolic steroids in doping control analysis. *Analytica Chimica Acta* **684**, 107-120 (2011).

43. Williamson, L.N., Zhang, G., Terry, A.V. & Bartlett, M.G. Comparison of Time-of-Flight Mass Spectrometry to Triple Quadrupole Tandem Mass Spectrometry for Quantitative Bioanalysis: Application to Antipsychotics.

Journal of Liquid Chromatography & Related Technologies **31**, 2737-2751 (2008).

44. McCloskey, D. A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. *Biotechnol Bioeng* **111**, 803-815 (2014).

45. Sambrook, J., and D. W. Russell Molecular cloning: a laboratory manual, 3rd ed., vol. A2.2. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001).

46. Fong, S.S. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

47. McCloskey, D., Utrilla, J., Naviaux, R., Palsson, B. & Feist, A. Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. *Metabolomics* **11**, 198-209 (2015).

48. Mashego, M.R. MIRACLE: mass isotopomer ratio analysis of U-13C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. *Biotechnol Bioeng* **85**, 620-628 (2004).

49. Bartha, A., Billiet, H.A.H., De Galan, L. & Vigh, G. Studies in reversed-phase ion-pair chromatography: III. The effect of counter ion concentration. *Journal of Chromatography A* **291**, 91-102 (1984).

50. Bartha, Á., Vigh, G., Billiet, H.A.H. & de Galan, L. Studies in reversed-phase ion-pair chromatography : IV. The rôle of the chain length of the pairing ion. *Journal of Chromatography A* **303**, 29-38 (1984).

51. Bartha, Á. & Vigh, G. Studies in reversed-phase ion-pair chromatography : V. Simultaneous effects of the eluent concentration of the inorganic counter ion and the surface concentration of the pairing ion. *Journal of Chromatography A* **395**, 503-509 (1987).

52. Iribarne, J.V. & Thomson, B.A. On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* **64**, 2287-2294 (1976).

53. Thomson, B.A. & Iribarne, J.V. Field induced ion evaporation from liquid surfaces at atmospheric pressure. *The Journal of Chemical Physics* **71**, 4451-4463 (1979).

54. Kebarle, P. & Tang, L. From ions in solution to ions in the gas phase - the mechanism of electrospray mass spectrometry. *Analytical Chemistry* **65**, 972A-986A (1993).

55. Tang, L. & Kebarle, P. Dependence of ion intensity in electrospray mass spectrometry on the concentration of the analytes in the electrosprayed solution. *Analytical Chemistry* **65**, 3654-3668 (1993).

56. Zhao, Y. Improved ruggedness of an ion-pairing liquid chromatography/tandem mass spectrometry assay for the quantitative analysis of the triphosphate metabolite of a nucleoside reverse transcriptase inhibitor in peripheral blood mononuclear cells. *Rapid Commun Mass Spectrom* **27**, 481-488 (2013).

57. Gao, S., Zhang, Z.-P. & Karnes, H.T. Sensitivity enhancement in liquid chromatography/atmospheric pressure ionization mass spectrometry using derivatization and mobile phase additives. *Journal of Chromatography B* **825**, 98-110 (2005).

58. Rayleigh, L. XX. On the equilibrium of liquid conducting masses charged with electricity. *Philosophical Magazine Series 5* **14**, 184-186 (1882).

59. Iavarone, A., Jurchen, J. & Williams, E. Effects of solvent on the maximum charge state and charge state distribution of protein ions produced by electrospray ionization. *Journal of the American Society for Mass Spectrometry* **11**, 976-985 (2000).

60. Cantwell, F.F. Retention model for ion-pair chromatography based on double-layer ionic adsorption and exchange. *J Pharm Biomed Anal* **2**, 153-164 (1984).

61. Bartha, A. & Vigh, G. Studies in reversed-phase ion-pair chromatography I. Adsorption isotherms of tetraalkylammonium ion-pair reagents on lichrosorb rp-18 in methanol-water eluents. *Journal of Chromatography A* **260**, 337-345 (1983).

62. Bartha, Á. & Vigh, G. Studies in ion-pair chromatography : II. Retention of positive and negative ions and neutral solutes in tetrabutylammonium bromide-containing methanol—water eluents on lichrosorb rp-18. *Journal of Chromatography A* **265**, 171-182 (1983).

[i] Increasing the concentration of TBA was also explored. Increasing the concentration above 15 mM resulted in noticeable ion-suppression, and changes in TBA concentration were no longer pursued.

CHAPTER 6:

MID Max: A LC-MS/MS method for measuring the precursor and product mass isotopomer distributions (MIDs) of metabolic intermediates and cofactors for metabolic flux analysis (MFA) applications

Abstract

The analytical challenges to acquire accurate isotopic data of intracellular metabolic intermediates for stationary, non-stationary, and dynamic metabolic flux analysis (MFA) are numerous. This work presents MID Max, a novel LC-MS/MS workflow, acquisition, and isotopomer deconvolution method for MFA that takes advantage of additional scan types that maximizes the number of mass isotopomer distributions (MIDs) that can be acquired in a given experiment. The analytical method was found to measure the MIDs of 97 metabolites, corresponding to 74 unique metabolite-fragment pairs (32 precursor spectra and 42 product spectra) with accuracy and precision. The compounds measured included metabolic intermediates in central carbohydrate metabolism and cofactors of peripheral metabolism (e.g., ATP). Using only a subset of the acquired MIDs, the method was found to improve the precision of flux estimations and number of resolved exchange fluxes for wild-type *E. coli* compared to traditional methods and previously published data sets.

Introduction

Metabolic labeling experiments involve the measurement of mass isotopomer distributions (MIDs) of metabolites at single or multiple time-points during dynamic or steady-state growth, or during isotopic dynamic or steady-state following the introduction of tracer into the culture medium[1]. GC-MS has traditionally been used to measure the MIDs of proteogenic amino acids derived from a single time-point taken at metabolic steady-state[2-6]. Proteogenic amino acids are stable, abundant, and have a carbon backbone that yields a multitude of informative fragmentations after electron impact ionization[2-6]. However, due to their slow turnover time, proteogenic amino acids are not suitable for capturing the transient isotopomer profiles of organisms that utilize a single-carbon containing compound (e.g., CO_2 in plants)[7], following a perturbation (e.g., after switching from one carbon source to another),[8] or during dynamic batch culture fermentation[9, 10]. Instead, intermediate metabolites with fast turnover that reflects the instantaneous biochemistry of the cell need to be measured. Metabolic intermediates in the cell are often at much lower abundance than proteogenic amino acids and are far less stable. Consequently, metabolic intermediates are difficult to accurately sample and extract[11, 12], and are mostly measured by LC-MS/MS due to its greater sensitivity and ability to resolve isomers of central carbohydrate metabolism, and softer ionization[13-16]. The combination of lower abundance, instability, and limited carbon backbone can result in less informative fragmentation patterns than proteogenic amino acids.

Recent studies looking at the MIDs of metabolic intermediates using LC-MS/MS and the ability to perform collision induced fragmentation have demonstrated the utility of measuring metabolic intermediates for metabolic labeling experiments[9, 17, 18]. The structural information provided by measuring multiple fragments following collision induced dissociation (CID) has been shown to increase the accuracy of flux calculations in metabolic labeling experiments when measuring intracellular intermediates[17, 19, 20, 20]. As an example, Ruhl et al, 2011 presented a compact method of isotopomer detection in order to conserve dwell time and maximize the number of fragment isotopomers that could be elucidated in a given run[18].

This work presents MID Max, a novel LC-MS/MS workflow (Figure 1), acquisition method, and isotopomer deconvolution method for MFA that takes advantage of additional scan types provided by hybrid instrumentation (Figure 2) to expand the number of MIDs that can be acquired in a given run. First, modifications to an existing sampling and extraction method are made to allow for an isotopic “snapshot” of the metabolic state. Second, the developed method is shown to accurately and precisely measure more MIDs of precursor and product fragments of intracellular central carbohydrate intermediates, biosynthetic intermediates, nucleotide phosphates, and cofactors compared to previous methods. Using only a subset of the acquired data, it is shown that an improvement in the precision of measured fluxes can be achieved over traditional approaches.

Material and Methods

Standards and reagents

Uniformly labeled ^{13}C glucose and 1-^{13}C glucose was purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA). Unlabeled glucose and other media components were purchased from Sigma-Aldrich (St. Louis, MO). LC-MS reagents were purchased from Honeywell Burdick & Jackson® (Muskegon, MI) and Sigma-Aldrich (St. Louis, MO).

Biological material

Replicate samples of *E. coli* K-12 MG1655 (ATCC 700926), obtained from the American Type Culture Collection (Manassas, VA), were grown in unlabeled or labeled glucose M9 minimal media[21] with trace elements[22] and sampled from a water bath that was maintained at 37 °C and aerated at 700 RPM during batch exponential growth in 500 mL Erlenmeyer flasks.

Extraction and sampling:

Samples were taken and extracted using a modified version of the fast Swinnex® filtration approach described previously[23]. The modifications made included the use of 47 mm filter and filter housing to accommodate the increased amount of culture broth and overall biomass sampled (10 mL at an OD600 of 1.0). Further details are described in the supplemental material.

Instrumentation

A XSELECT HSS XP 150 mm × 2.1 mm × 2.5 µm (Waters®, Milford, MA) with a Prominence UFLC XR HPLC (Shimadzu, Columbia, MD) was used for chromatographic separation. An AB SCIEX QTRAP® 5500 mass spectrometer

(AB SCIEX, Framingham, MA) operated in negative mode was used for detection. Further details of the LC-MS/MS system and gradient used were described previously[16]. The list of MRM transitions is provided in supplemental table S-3. The AB SCIEX acquisition files for the method are available upon request. Detailed protocols for running the method can also be made available upon request.

Raw data processing:

MRM data was integrated using multiQuant® 3.0.1. Product ions corresponding to each MRM transition were extraction using PeakView® 2.2. The quantitation methods for MultiQuant® and the precursor ion annotation method for PeakView® can be made available upon request. The detailed protocols for using the software along with the methods can also be made available upon request.

Product ion spectral library generation:

Product ion spectra were acquired for unlabeled standards by direct injection. Unlabeled standards were diluted to an appropriate concentration for the detector in a 50/50 (v/v%) of water + 0.1% formic and Acetonitrile + 0.1% formic acid. The syringe speed was 7 uL/min. Source parameters are given in Supplemental Table S . Parameters for declustering potential (DP), entrance potential (EP), collision energy (CE), and collision cell exit potential (CXP) were optimized for each metabolite prior to spectral acquisition. Further details are described in the supplemental material.

Calculation of isotopomers

Multiple reaction monitoring scans (MRMs) acquired from samples collected in triplicate and analyzed in duplicate (n=6) were integrated using MultiQuant® 3.0.1. Mass spectras from EPI and ER scans for all samples analyzed were extracted using peakView® 2.2. Precursor and product MIDs from MRM acquisitions were calculated using in-house scripts. Mass spectras from EPI and ER scans were processed using in-house scripts. Theoretical isotopomer spectra for each compound were calculated using open-source python modules. The data processing scripts are hosted on Github (https://github.com/dmccloskey/MDV_utilities).

Metabolic Flux Analysis

The *E. coli* model used for MFA included iRL2013[24]. MFA simulations were conducted with MATLAB® and INCA v1.3[25]. Confidence intervals observable fluxes were calculated using published methods[26]. Acquired metabolite MIDs for Acetyl-CoA, AMP, ATP, FAD, and UMP were not included in the flux estimations. See supplemental information for additional details.

Results and Discussion

Sampling and extraction

A major challenge to stationary, non-stationary, and dynamic MFA experiments aimed at measuring the isotopic distribution of intracellular metabolites is to preserve an isotopic "snapshot" of metabolism at the time of sampling. This is a non-trivial task[11] given the rapid turnover, cell leakage, and general instability of intracellular metabolites. It is made more problematic by the increased biomass needed to acquire the complete MID for a given compound. This problem was resolved by sampling and extracting samples using a modified version of the fast Swinnex® filtration approach described previously[23]. The modifications made included the use of 47 mm filter and filter housing to accommodate the increased amount of culture broth and overall biomass sampled (10 mL at an OD600 of 1.0) without compromising the speed at which cultures were previously found to be sampled (Supplemental Figure 1).

Information-dependent acquisition (IDA) method development

A primary analytical goal of metabolic labeling experiments is the accurate acquisition of as many mass isotopomer distributions (MIDs) as possible to provide more data points with which to constrain the solution space when calculating flux ratios or absolute fluxes[27]. MIDs for compounds can be obtained through either an enhanced resolution (ER, Supplemental Figure 2) scan of the precursor ion or a series of multiple MRM scans (Figure 3a and b) on the instrument used in this study. Both scan types were compared when implemented on the QTRAP instrument at a nominal concentration of biomass. It

was found that an ER scan was not able to accurately measure as many theoretical MIDs in a biological matrix as multiple MRM scans (Table S-1 and supplemental discussion). Therefore, a general scheme to employ multiple MRM scans was selected that maximized signal intensity and minimized the number of MRM transitions needed to capture the entire theoretical precursor isotopic spectrum.

An information-dependent acquisition (IDA) method was then constructed whereby an enhanced product ion (EPI) scan was triggered for any of the MRM transitions that corresponded to an isotopomer state for a given compound (Figure 2 and Supplemental Figure 3). The IDA-EPI scan required less dwell time than using multiple MRMs that correspond to each of the different product fragments. This allowed for greater sensitivity and conserved cycle time for additional compounds. It was found that several masses corresponding to different isotopic states for several fragments could often be acquired for each EPI scan. By normalizing each product ion spectra acquired per EPI scan to the contribution of the precursor isotopomer determined by the MRM scan, the product-ion spectra could be determined (Figure 3c).

Product ion spectral library generation

A library of product ion spectra was generated from the injection of pure standards in order to deduce the structural basis for the observed product ion peaks (Figure 4). Structural annotation was first based on data from the literature[14-16, 18, 28] and then from simulated theoretical fragmentation that had minimal mass error between the predicted and observed ion mass. Simulated

product ion peaks that could be derived from multiple fragments with minimal mass error were either omitted or compared against the expected isotopomer spectra as described below. All 240 product ion spectral files and 124 annotations corresponding to 167 compounds that were generated during this process are provided as supplemental material (Supplemental Table S-12, Supplemental Files S-1, S-2, S-3). Carbon positions of each fragment structure were assigned indexes based on historical mapping (e.g., carbon mapping of glucose in traditional MFA models) or assigned indexes based on the final carbon mapping network for those metabolites that were not measured nor included in traditional MFA models (Table S-4).

Analytical method validation

An extensive screening procedure was initially conducted to narrow down the number of compounds that would be included in the acquisition method. A set of target metabolites that were consistently found to show good reproducibility, high sensitivity, and relatively large abundance in most biological samples analyzed were screened for use in the fluxomics acquisition method from a pool of over 100 compounds validated for the LC-MS/MS assay previously[16]. Compounds with similar retention times that had the potential for overlap in their isotopomer transitions were removed. Compounds with or without isotopic overlap were first confirmed or disconfirmed by running groups of pooled standards as well as concentrated pools of unlabeled *E. coli* biomass. Second, compounds with or without isotopic overlap were checked when running samples of labeled ^{13}C *E. coli* biomass. The use of uniformly labeled *E. coli*

biomass also served to identify compounds with ghost or non-sense peaks in any of the isotopomer transitions that would interfere with analysis. Compounds that were removed due to isotopic overlap included many of the organic acids and a majority of the nucleotide mono-, di-, and tri-phosphates. Additionally, many of the amino acids and nucleotides and bases were screened out due to scan rate considerations (i.e., dwell time) early in the chromatographic gradient. Compounds that were screened are listed in Supplemental Table S-11. Those compounds that remained were further screened by running samples of unlabeled *E. coli* as described below.

The ability of MID Max to reproduce the theoretical isotopic spectrum for samples of unlabeled *E. coli* biomass was tested[5, 29]. The method was able to capture the theoretical isotopomer spectrum for 97 MIDs consisting of 74 unique metabolite-fragment pairs (32 precursor spectra and 42 product spectra) (Table 1 and 2). The MIDs included central carbohydrate intermediates, biosynthetic pathway intermediates, as well as cofactors and nucleotide phosphates, with less than 1% average absolute deviation from the theoretical (Table S-2). Not surprisingly, it was found that the average absolute deviation for the fragment isotopomers determined only from the MRM scan (0.69%) is slightly lower than those determined by the EPI scans (1.09%). This is due to the fact that errors in the MRM scan are propagated by the errors of the EPI scan when individual EPI scans are normalized to the contribution of the precursor isotope from the MRM scan. However, for most metabolite-fragment pairs, this additional source of error was found to be minimal.

In addition, 29 MIDs were acquired with a higher (4.8%) average absolute deviation from the theoretical (Table 2). The higher deviation from the theoretical is due to insufficient resolution to capture the full unlabeled MID spectrum. It may be possible to capture the full MID spectrum on a different tracer (see below). This indicates that the method may be able to acquire 126 MIDs in a single run. This is more MIDs than what has been previously reported[30].

Improved precision of measured flux values

The ability of the method developed here to acquire precise MIDs of peripheral metabolites in practice was tested. Using MID Max, mass spectra acquired from wild-type *E. coli* grown on an 80/20 mixture of 1-^{13}C/U-^{13}C, and sampled and extracted during exponential growth were analyzed. This tracer scheme is commonly used in MFA studies for its ability to successfully resolve a multitude of paths [2, 3]. The method was able to capture the precursor MIDs of peripheral metabolites and several fragments with acceptable precision (see Figure 5 for an example, and Supplemental Table S8). Specifically, it was able to measure the MIDs for the precursor and multiple fragments for the five key periphery metabolites, Acetyl-CoA, AMP, ATP, FAD, and UMP, with an average relative standard deviation (RSD) of 18.4%. This result shows promise that these metabolites could be included in MFA modeling.

The accuracy and precision of fluxes estimated through an MFA application was tested using measured metabolite MIDs of central carbohydrate metabolism (62 of 87 total fragments) acquired using MID Max. A previously validated MFA model[24] was used in the analysis. First, the effect of the

additional product ion scans on the precision of observable net fluxes was tested by including varying amounts of MIDs generated by the product ion scans. An increase in estimated average net flux precision (from 0.056 to 0.050) was found as additional product ion scans were included in the estimation procedure (Table 3). Second, accuracy was tested by comparing flux values estimated using the acquired data to previously published results[24, 31] using the same organism and MFA model during steady-state growth on glucose, but with GC-MS derived MIDs. Flux estimations were found to agree with the previously published values[24, 31] (Table 4). An average absolute deviation of 10.8% (flux normalized to glucose uptake, Supplemental Table S10) for all tracer schemes used in the study was found for all reactions not including the exchange of unlabeled carbon dioxide. It should be noted that greater than 94% of the deviation (i.e., of the 10.8%) from the published values could be attributed to the reactions corresponding to ATP maintenance, oxidative phosphorylation, and acetate exchange. When those reactions were excluded, an average absolute deviation of 6.0% was found (Table 4). Third, the precision of estimated flux values was compared using the same data sets. It was found that the MIDs acquired using the method described here were able to estimate the same or a greater number of reactions with better precision for all of the tracer schemes tested in the published study (Table 4 and supplemental Table S10). This included the most similar tracer ([1] + [U]Glc (4:1)) to the 80/20 mixture of 1-^{13}C/U-^{13}C used here. In addition, the method was also found to resolve a greater number of exchange fluxes than any of the individual tracers used in the published study (Table 4).

This most likely directly relates to the number of MIDs that MID Max is able to acquire as well as the ability to directly measure intracellular intermediates and their fragments.

Conclusion:

MID Max, a LC-MS/MS acquisition method and isotope recapitulation algorithm was described that utilizes hybrid instrumentation and multiple scan types to increase the number of MIDs that can be measured in a single run. A library of product ion scans and spectral annotation was generated in the process. This is in itself a valuable resource to the scientific community. The method was validated through a series of control experiments using pooled standards, unlabeled *E. coli* biomass, and uniformly labeled ^{13}C *E. coli* biomass. It was shown that even when using a subset of the acquired data, flux precision can be improved and the number of resolved exchange fluxes increased. In order to utilize the full data-set, a genome-scale MFA model will be needed. The analytical workflow and methodology presented is scalable to improvements in instrumentation over what is currently available, can be applied to MFA experiments using tracer schemes other than ^{13}C (e.g., deuterium [32]), and presents advances over current analytical methods that can be used in stationary, non-stationary, or dynamic MFA experiments.

Acknowledgments

Chapter 6, in full, is a reformatted reprint of “MID Max: LC–MS/MS Method for Measuring the Precursor and Product Mass Isotopomer Distributions of Metabolic Intermediates and Cofactors for Metabolic Flux Analysis Applications.” McCloskey D, Young JD, Xu S, Palsson BO, Feist AM. Anal Chem. 2016 Jan 19;88(2):1362-70. doi: 10.1021/acs.analchem.5b03887. The dissertation/thesis author was the primary investigator and author of the paper.

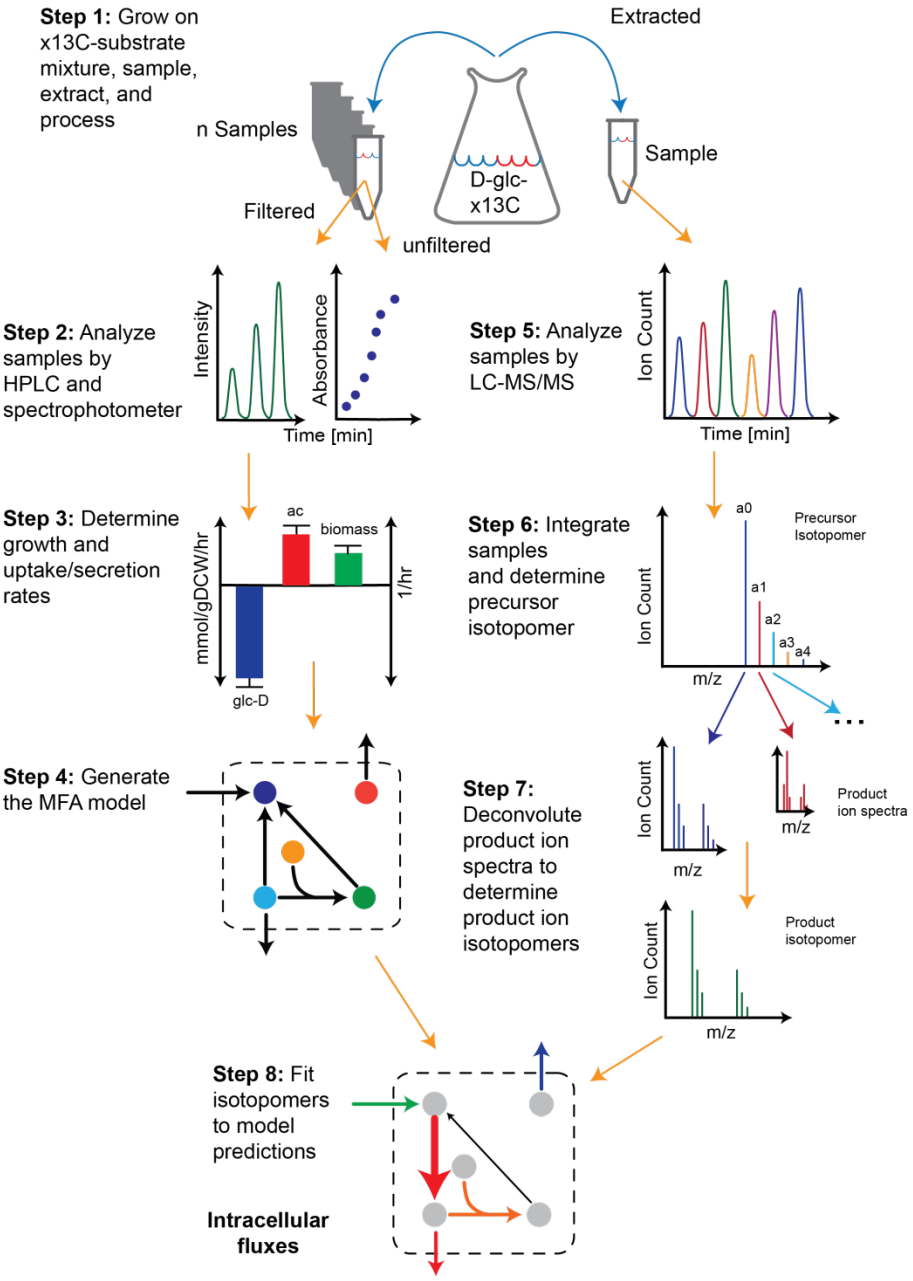

Figure 6.1: Overview of the LC-MS-enabled fluxomics experiment using hybrid instrumentation. Cultures grown on a single or a combination of heavy isotope tracers are rapidly sampled and extracted. In parallel, or during a separate experiment, filtrate samples and biomass samples are taken periodically during steady-state growth. Extracted samples are further processed and then metabolites are separated and acquired on the LC-MS instrument. After integration of peaks for each sample and processing of acquired product ion spectra, precursor and product MIDs are calculated. Filtrate samples are analyzed by HPLC to determine substrate uptake and secretion rates. Biomass samples are analyzed by a spectrophotometer to determine the growth rate. The MIDs, substrate uptake and secretion rates, and growth rate are fitted to a reaction network of metabolism. The estimated fluxes that minimize the error between the predicted and measured values are then calculated.

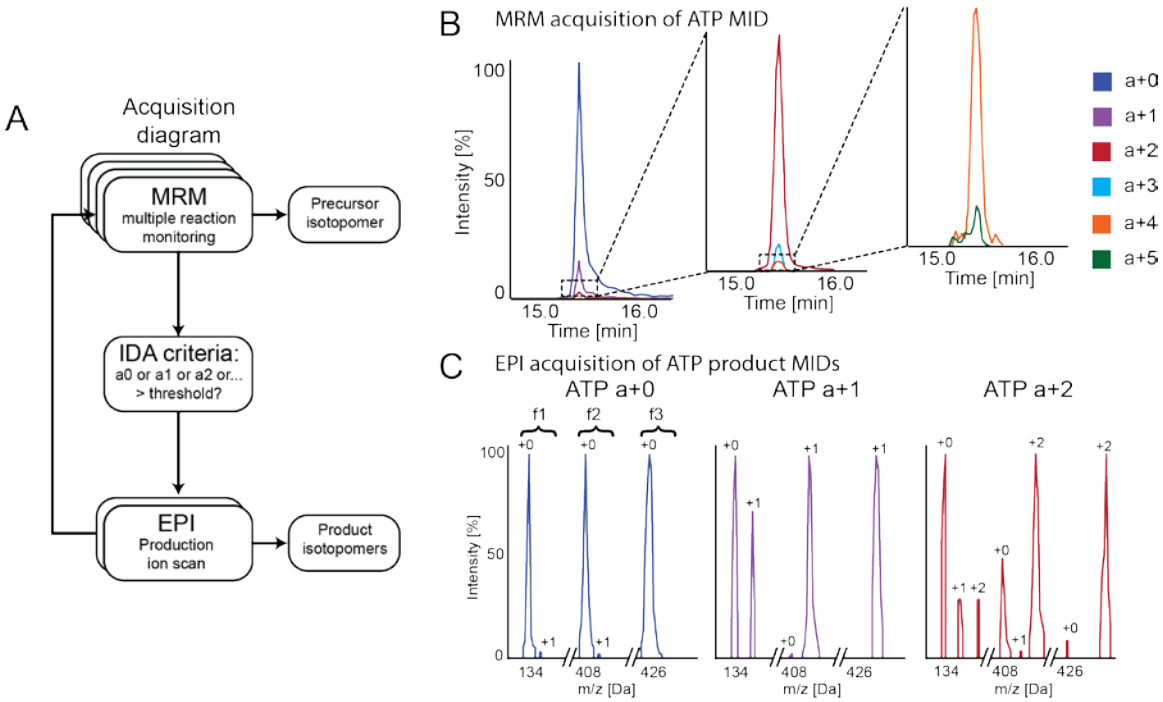

Figure 6.2: LC-MS/MS acquisition method and example. A) Acquisition Diagram. When an MRM transition exceeds a specified threshold, an IDA is triggered that utilizes an enhanced product ion (EPI) scan. The spectra of the EPI scans are then used to infer additional product isotopomers not captured by the MRM scan. B) MRM acquisition of the ATP isotopomer spectra in an *E. coli* matrix using the metabolic labeling method. Individual MRM transitions corresponding to isotopomers of ATP (a+0, a+1, …) were acquired. C) EPI acquisition of the ATP product isotopomers. The product ion scans that reveal the labeling states of product ions (f1+0, f1+1, …, f2+0, f2+1, …) triggered by each MRM transition corresponding to an isotopomer of ATP are extracted, processed, and used to determine the product isotopomers.

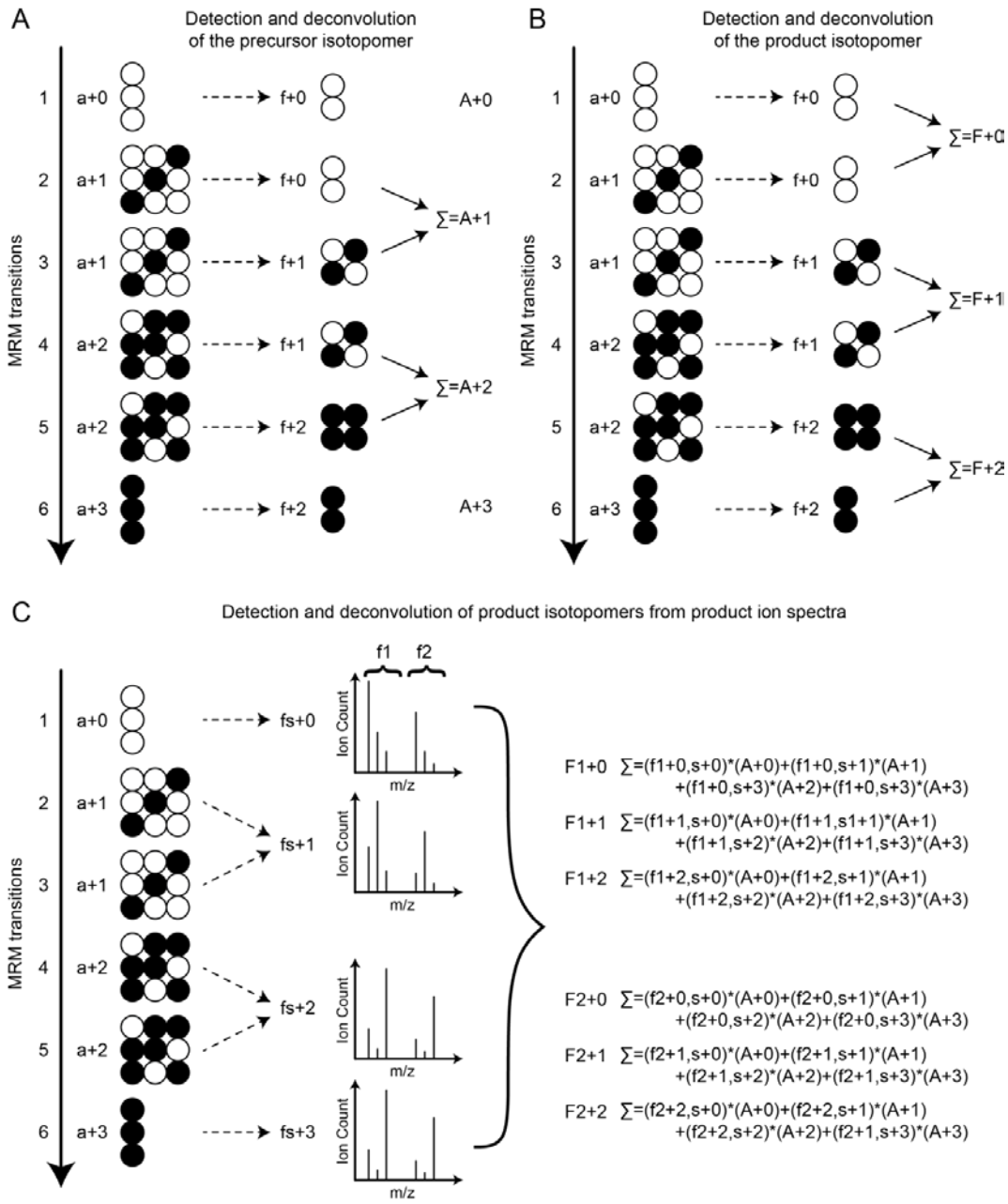

Figure 6.3: Detection and deconvolution of precursor and product isotopomers. A) A toy example of MRM transitions corresponding to a precursor with 3 carbons and a product with 2 carbons is shown. The precursor isotopomers (A+0, A+1, …) were calculated by summing over the intensities of the transitions measured for each combination of product isotopomer that contributed to the a+0, a+1, … mass. B) The product isotopomers (F+0, F+1,…) were calculated by summing over the intensities of transitions measured for each combination of precursor isotopomer that contributed to the f+0, f+1, … mass. C) Additional product isotopomers (F1+0,F1+1,…,F2+0,F2+1,…) were calculated from the product ion scans (fs+0, fs+1, …) triggered by MRM scans corresponding to a precursor isotopomer (a+0, a+1, …). Each additional product isotopomer was calculated by summing over the intensities of each isotopomer in each spectrum (f1+0,s+0, f1+0,s+1, …, f1+1,s+0, f1+1,s+1, …, and f2+0,s+0, f2+0,s+1, …, f2+1,s+0, f2+1, …) and normalizing to the corresponding precursor isotopomer (A+0, A+1, …) determined previously.

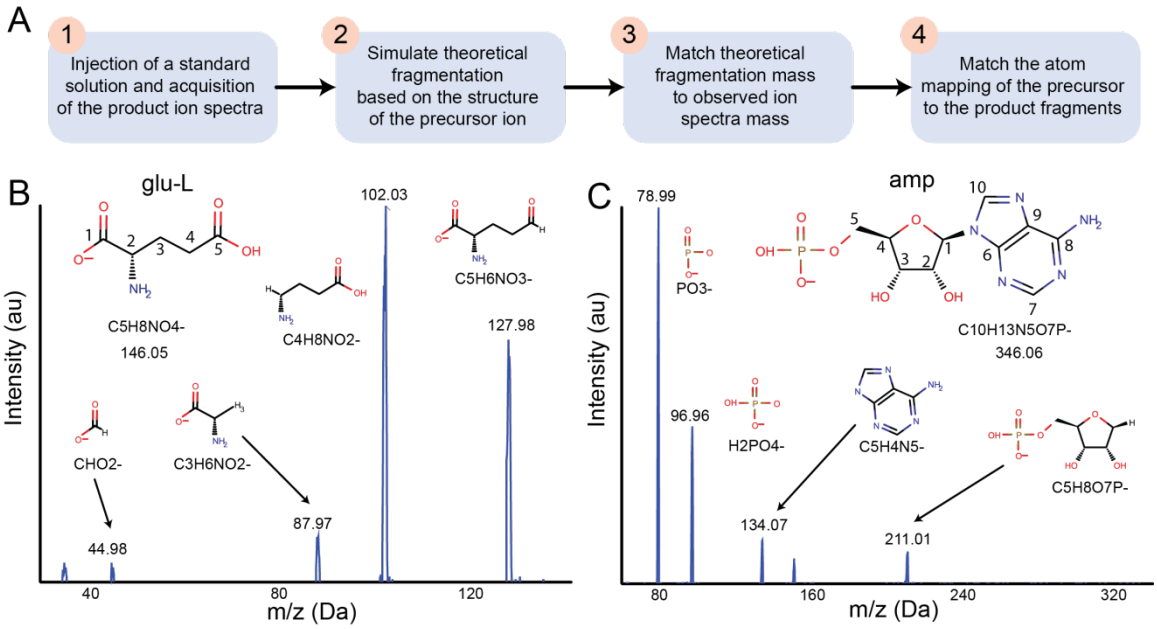

Figure 6.4: Product ion spectral library generation and fragment annotation. A) Workflow used to map carbon positions to product ion spectra from the direct injection of pure, unlabeled standards. Panels B and C correspond to examples using the workflow for glutamate (glu-L) and AMP (amp), respectively.

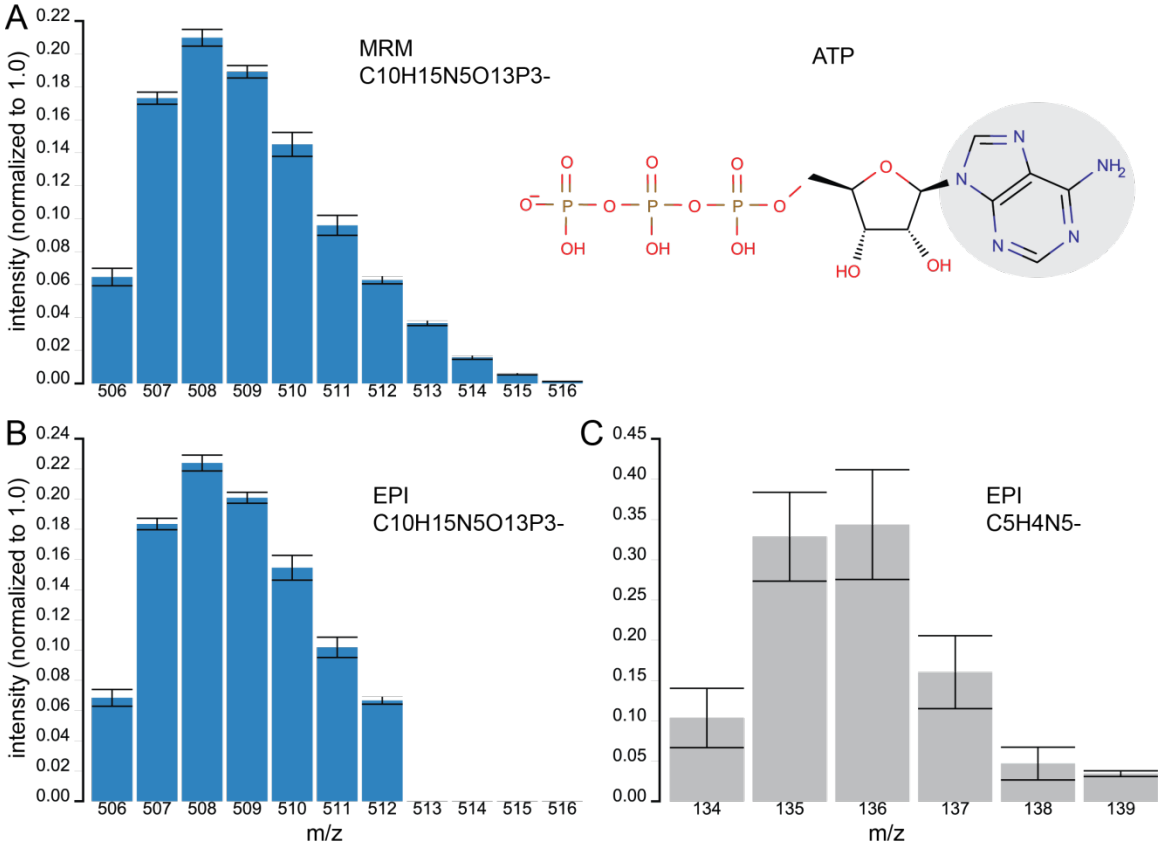

Figure 6.5: Measured MIDs of ATP measured from wild-type *E. coli grown* on an 80/20 mixture of 1-^{13}C/U-^{13}C**.** The structure of ATP is shown in the top right. A) The calculated precursor MID calculated from MRMs. B) The calculated precursor MID calculated from EPI scan. C) The calculated product MID calculated from EPI scan.

Table 6.1: List of validated metabolites and fragments:

Met ID	Metabolite Name	Fragment Formula
23dpg	3-Phospho-D-glyceroyl phosphate	C3H7O10P2-
6pgc	6-Phospho-D-gluconate	C6H12O10P-
6pgc	6-Phospho-D-gluconate	C6H10O9P-
6pgc	6-Phospho-D-gluconate	C5H12O7P-
6pgc	6-Phospho-D-gluconate	C5H12O8P-
Accoa	Acetyl-CoA	C23H37N7O17P3S-
accoa	Acetyl-CoA	C23H36N7O14P2S-
accoa	Acetyl-CoA	C13H23N2O10P2S-
accoa	Acetyl-CoA	C10H14N5O10P2-
accoa	Acetyl-CoA	C10H12N5O9P2-
accoa	Acetyl-CoA	C10H11N5O6P-
acon-C	Aconitate-C	C6H5O6-
acon-C	Aconitate-C	C6H3O5-
acon-C	Aconitate-C	C5H5O4-
akg	Alpha-ketoglutarate	C5H5O5-
akg	Alpha-ketoglutarate	C4H5O3-
akg	Alpha-ketoglutarate	C2HO3-
amp	AMP	C5H4N5-
amp	AMP	C10H13N5O7P-
asp-L	L-aspartate	C4H6NO4-
asp-L	L-aspartate	C4H6NO2-
asp-L	L-aspartate	C4H3O4-
asp-L	L-aspartate	C3H6NO2-
atp	ATP	C10H15N5O13P3-
atp	ATP	C10H14N5O10P2-
atp	ATP	C10H12N5O9P2-
atp	ATP	C10H11N5O6P-
dhap	Dihydroxyacetone phosphate	C3H6O6P-
fad	FAD	C27H32N9O15P2-
fad	FAD	C17H18N4O8P-
fad	FAD	C10H13N5O7P-
fdp	D-Fructose 1,6-bisphosphate	C6H13O12P2-
fdp	D-Fructose 1,6-bisphosphate	C6H10O8P-
fdp	D-Fructose 1,6-bisphosphate	C4H8O6P-
fdp	D-Fructose 1,6-bisphosphate	C6H8O7P-
g1p	D-Glucose 1-phosphate	C6H12O9P-
g6p	D-Glucose 6-phosphate	C6H12O9P-
g6p	D-Glucose 6-phosphate	C4H8O7P-
glu-L	L-glutamate	C5H8NO4-
glu-L	L-glutamate	C5H6NO3-
glu-L	L-glutamate	C4H8NO2-
glyc3p	Glycerol 3-phosphate	C3H8O6P-
glyclt	Glycolate	CH3O2-
glyclt	Glycolate	C2H3O3-
icit	Isocitrate	C6H7O7-
icit	Isocitrate	C6H5O6-
icit	Isocitrate	C5H3O3-
mal-L	Malate	C4H5O5-
mal-L	Malate	C4H3O4-
met-L	L-Methionine	CH3S-
met-L	L-Methionine	C5H10NO2S-
pep	Phosphoenolpyruvate	C3H4O6P-
phe-L	L-phenylalanine	C9H7O2-
phe-L	L-phenylalanine	C9H10NO2-
phpyr	Phenylpyruvate	C9H7O3-
phpyr	Phenylpyruvate	C8H7O
phpyr	Phenylpyruvate	C7H7-
Pool_2pg_3pg		C3H6O7P-
Pool_2pg_3pg	Pool of D-Glycerate 2-phosphate and 3-Phospho-D-glycerate	C2H6O5P-
prpp	5-Phospho-alpha-D-ribose 1-diphosphate	C5H9O10P2-
prpp	5-Phospho-alpha-D-ribose 1-diphosphate	C5H12O14P3-
pyr	Pyruvate	C3H3O3-
r5p	D-Ribose 5-phosphate	C5H10O8P-
ru5p-D	D-Ribulose 5-phosphate	C5H10O8P-
s7p	Sedoheptulose 7-phosphate	C7H14O10P-
skm	Shikimate	C7H9O5-
skm	Shikimate	C6H5O-
succ	Succinate	C4H5O4-
succ	Succinate	C4H3O3-
thr-L	L-Threonine	C4H8NO3-
thr-L	L-Threonine	C2H4NO2-
ump	UMP	C9H9N2O6-
ump	UMP	C9H12N2O9P-
ump	UMP	C4H3N2O2-

Table 6.2: Summary of the acquisition method validation on unlabeled *E. coli* biomass.

	Metabolite-fragment pairs	**Absolute deviation from theoretical %**	**RSD %**
MRM + EPI	97	0.90	9.49
MRM	45	0.69	7.52
EPI	52	1.09	11.19
Partial	29	4.78	6.90

Table 6.3: MFA Model and net flux estimation statistics using a subset of the acquired MIDs.

model	fitted fluxes	fitted fragments	fitted dof	fitted chi2	chi2 pass	observable fluxes	fluxes	average observable flux precision
iRL2013 (MRM + EPI)	3	62	224	243.4	TRUE	56	75	0.050
iRL2013 (MRM)	3	36	119	134.6	TRUE	57	75	0.056
iRL2013 (unique MRM + EPI)	3	47	164	186.3	TRUE	55	75	0.054

Table 6.4: Accuracy and precision using a subset of the acquired MIDs (i.e., iRL2013 (MRM+EPI)) compared to previously published results using iRL2013 and GC-MS derived MIDs[31]. *Reactions corresponding to the exchange of unlabeled carbon dioxide, ATP maintenance, oxidative phosphorylation, and acetate exchange were not included in the calculations shown. Values considering all reactions are shown in Supplemental Table S-10.

Tracer	average absolute deviation of estimated flux values (% flux normalized to glucose uptake)*	number of net fluxes with improved precision (65 total)*	increased number of resolved exchanged fluxes (21 total)
[1,2]Glc	4.53	9	7
[2,3]Glc	4.50	23	8
[4,5,6]Glc	4.89	15	8
[2,3,4,5,6]Glc	5.66	19	7
[1] + [4,5,6]Glc (1:1)	4.72	21	8
[1] + [U]Glc (1:1)	4.91	21	8
[1] + [U]Glc (4:1)	5.16	27	7
20% [U]Glc	4.74	35	8
[1]Glc	6.29	29	8
[2]Glc	5.00	15	6
[3]Glc	6.23	33	8
[4]Glc	17.48	39	10
[5]Glc	5.14	5	6
[6]Glc	5.63	13	9

References:

1. Jazmin, L. & Young, J. in Systems Metabolic Engineering, Vol. 985. (ed. H.S. Alper) 367-390 (Humana Press, 2013).

2. Fischer, E., Zamboni, N. & Sauer, U. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Anal Biochem* **325**, 308-316 (2004).

3. Fischer, E. & Sauer, U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nat Genet* **37**, 636-640 (2005).

4. Zamboni, N., Fischer, E. & Sauer, U. FiatFlux--a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* **6**, 209 (2005).

5. Zamboni, N., Fendt, S.M., Ruhl, M. & Sauer, U. (13)C-based metabolic flux analysis. *Nat Protoc* **4**, 878-892 (2009).

6. Schellenberger, J. Predicting outcomes of steady-state (1)(3)C isotope tracing experiments using Monte Carlo sampling. *BMC Syst Biol* **6**, 9 (2012).

7. Ma, F., Jazmin, L.J., Young, J.D. & Allen, D.K. Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. *Proceedings of the National Academy of Sciences* **111**, 16967-16972 (2014).

8. Yuan, J., Bennett, B.D. & Rabinowitz, J.D. Kinetic flux profiling for quantitation of cellular metabolic fluxes. *Nat Protoc* **3**, 1328-1340 (2008).

9. Toya, Y. 13C-metabolic flux analysis for batch culture of Escherichia coli and its Pyk and Pgi gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. *Biotechnol Prog* **26**, 975-992 (2010).

10. Toya, Y. Direct measurement of isotopomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis. *J Chromatogr A* **1159**, 134-141 (2007).

11. Millard, P., Massou, S., Wittmann, C., Portais, J.C. & Létisse, F. Sampling of intracellular metabolites for stationary and non-stationary 13C metabolic flux analysis in Escherichia coli. *Analytical Biochemistry* **465**, 38-49 (2014).

12. McCloskey, D., Utrilla, J., Naviaux, R., Palsson, B. & Feist, A. Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. *Metabolomics* **11**, 198-209 (2015).

13. Lu, W. Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal Chem* **82**, 3212-3221 (2010).

14. Buescher, J.M., Moco, S., Sauer, U. & Zamboni, N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. *Anal Chem* **82**, 4403-4412 (2010).

15. Luo, B., Groenke, K., Takors, R., Wandrey, C. & Oldiges, M. Simultaneous determination of multiple intracellular metabolites in glycolysis, pentose phosphate pathway and tricarboxylic acid cycle by liquid chromatography-mass spectrometry. *J Chromatogr A* **1147**, 153-164 (2007).

16. McCloskey, D., Gangoiti, J., Palsson, B. & Feist, A. A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites. *Metabolomics*, 1-13 (2015).

17. Choi, J. & Antoniewicz, M.R. Tandem mass spectrometry: a novel approach for metabolic flux analysis. *Metab Eng* **13**, 225-233 (2011).

18. Rühl, M. Collisional fragmentation of central carbon metabolites in LC-MS/MS increases precision of 13C metabolic flux analysis. *Biotechnology and Bioengineering* **109**, 763-771 (2012).

19. Ruhl, M. Collisional fragmentation of central carbon metabolites in LC-MS/MS increases precision of (1)(3)C metabolic flux analysis. *Biotechnol Bioeng* **109**, 763-771 (2012).

20. Tepper, N. & Shlomi, T. Efficient Modeling of MS/MS Data for Metabolic Flux Analysis. *PLoS One* **10**, e0130213 (2015).

21. Sambrook, J., and D. W. Russell Molecular cloning: a laboratory manual, 3rd ed., vol. A2.2. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001).

22. Fong, S.S. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

23. McCloskey, D., Utrilla, J., Naviaux, R., Palsson, B. & Feist, A. Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. *Metabolomics*, 1-12 (2014).

24. Leighty, R.W. & Antoniewicz, M.R. COMPLETE-MFA: complementary parallel labeling experiments technique for metabolic flux analysis. *Metab Eng* **20**, 49-55 (2013).

25. Young, J.D. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* **30**, 1333-1335 (2014).

26. Antoniewicz, M.R., Kelleher, J.K. & Stephanopoulos, G. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metab Eng* **8**, 324-337 (2006).

27. Chang, Y., Suthers, P.F. & Maranas, C.D. Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. *Biotechnol Bioeng* **100**, 1039-1049 (2008).

28. Bajad, S.U. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J Chromatogr A* **1125**, 76-88 (2006).

29. Antoniewicz, M.R., Kelleher, J.K. & Stephanopoulos, G. Accurate assessment of amino acid mass isotopomer distributions for metabolic flux analysis. *Anal Chem* **79**, 7554-7559 (2007).

30. Huege, J. GC-EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after 13CO2 labelling. *Phytochemistry* **68**, 2258-2272 (2007).

31. Crown, S.B., Long, C.P. & Antoniewicz, M.R. Integrated 13C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. *Metab Eng* **28**, 151-158 (2015).

32. Lewis, C.A. Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Mol Cell* **55**, 253-263 (2014).

CHAPTER 7:

Modeling Method for Increased Precision and Scope of Directly Measurable Fluxes at a Genome-Scale

Abstract:

Metabolic flux analysis (MFA) is considered to be the gold standard for determining the intracellular flux distribution of biological systems. The majority of work using MFA has been limited to core models of metabolism due to challenges in implementing genome-scale MFA and the undesirable trade-off between increased scope and decreased precision in flux estimations. This work presents a tunable workflow for expanding the scope of MFA to the genome-scale without trade-offs in flux precision. The genome-scale MFA model presented here, iDM2014, accounts for 537 net reactions, which includes the core pathways of traditional MFA models and also covers the additional pathways of purine, pyrimidine, isoprenoid, methionine, riboflavin, coenzyme A, and folate, as well as other biosynthetic pathways. When evaluating the iDM2014 using a set of measured intracellular intermediate and cofactor mass isotopomer distributions (MIDs)[1], it was found that a total of 232 net fluxes of central and peripheral metabolism could be resolved in the *E. coli* network. The increase in scope was shown to cover the full biosynthetic route to an expanded set of bioproduction pathways, which should facilitate applications such as the design of more complex bioprocessing strains and aid in identifying new antimicrobials. Importantly, it was found that there was no loss in precision of core fluxes when

compared to a traditional core model, and additionally there was an overall increase in precision when considering all observable reactions.

Introduction

The intracellular flux distribution of metabolites is considered by many to define the functional state of the cell. A precise readout of the flux profile of an organism is important for discovering and refining our knowledge of the metabolic capabilities of organisms, [2-9] as well as to engineer organisms to more efficiently drive precursor metabolites towards product metabolites for the sustainable production of commodity chemicals and biotherapeutics [10-17]. With limited experimental data, methods exist to predict the intracellular flux of the cell using genome-scale metabolic models (i.e., “M models”) (see [18] and [19] for a review). Furthermore, it has been shown that by incorporating the expression network with the metabolic network (i.e., “ME model”), more accurate predictions of intracellular fluxes can be achieved as more constraints are added [20-22]. Even with these *in silico* constraint-based modeling approaches, the gold standard for assessing the intracellular flux distribution of the cell is via metabolic labeling experiments [23, 24].

Once mass isotopomer distributions (MIDs) of labeled cellular components, either proteogenic amino acids [24-28] or central carbon intracellular metabolites [2, 29],have been determined, intracellular flux distributions can then be calculated using metabolic flux analysis (MFA). A prerequisite of MFA is a valid stoichiometric model of the organism’s metabolism and a valid mapping of the elements that comprise the labeled tracer through the network. Core models that account for the minimal amount of carbon metabolism have traditionally been used in MFA calculations [8, 30]. These core models have fewer free fluxes, which

provides greater degrees of freedom and allows for greater statistical confidence in the MFA fitting procedure. Additionally, core models have a fast simulation time and an atom mapping network that is readily available in the literature [30]. However, minimal core models do not account for cofactor usage, biosynthesis, or salvage, and thus do not provide flux information for many pathways that are important for understanding and engineering the physiology of the cell. While genome-scale atom mappings have emerged for various organisms [31, 32], and a recent study using a genome-scale model with MFA has emerged[33], genome-scale models have not been widely adopted in MFA. Several reasons for this include slower simulation times, a lack of available isotope measurement of metabolites outside proteogenic amino acids and central carbon intermediates, the nontrivial integration and reconciliation of genome-scale atom mapping models with genome-scale stoichiometric models, and a loss in flux precision due to the increase number of alternate biosynthetic routes[33].

Previous work to integrate ^{13}C MFA with genome-scale modeling has been conducted. These attempts have used ^{13}C MFA data to constrain core metabolic pathways while using FBA to solve the full genome-scale model or have included ^{13}C MFA data as additional constraints in the FBA problem [34-36]. Given that these approaches do not include a genome-scale carbon mapping network, the methods essentially extrapolate information from metabolic labeling experiments to the genome-scale instead of directly calculating the flux values based off of the MIDs themselves. This approach can be problematic as variability in labeling patterns due to the recycling of metabolic intermediates

from peripheral metabolic pathways is not captured[33]. In addition, information obtained from the labeling of cofactors and other peripheral metabolites cannot be included in the analysis. A more recent effort has demonstrated the ability to calculate fluxes at a genome-scale within the MFA framework[33]. Irrespective of labeling substrate, the authors found that an inherent disadvantage of ^{13}C MFA at the genome-scale is a loss in flux precision due to the increased number of alternate biosynthetic routes and carbon recycling from peripheral metabolism[33].

This work expands the scope of MFA to the genome-scale without a loss in flux precision. A genome-scale model based on the most recent *E. coli* reconstruction[37] with a complete carbon mapping was constructed and compared to MFA models that account for varying degrees of cofactor usage, peripheral metabolism, and biomass composition. It was found that the increased scope gained when utilizing genome-scale models for MFA simulation does not come at a cost of loss in flux precision. It was shown that the precision of flux estimates using the genome-scale model can be improved through measurements of cofactor and peripheral metabolite MIDs.

Materials and methods

Standards and Reagents:

Uniformly labeled ^{13}C glucose and 1-^{13}C glucose was purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA). Unlabeled glucose and other media components were purchased from Sigma-Aldrich (St. Louis, MO). LC-MS reagents were purchased from Honeywell Burdick & Jackson® (Muskegon, MI) and Sigma-Aldrich (St. Louis, MO).

Biological Material:

Replicate samples of *E. coli* K-12 MG1655 (ATCC 700926), obtained from the American Type Culture Collection (Manassas, VA), were grown in unlabeled or labeled glucose M9 minimal media[38] with trace elements[39] and sampled from a water bath that was maintained at 37 °C and aerated at 700 RPM during batch exponential growth in 500 mL Erlenmeyer flasks. Samples were taken and extracted using a modified version of the fast Swinnex® filtration approach described previously[1, 40].

Instrumentation and data processing:

A XSELECT HSS XP 150 mm × 2.1 mm × 2.5 µm (Waters®, Milford, MA) with a Prominence UFLC XR HPLC (Shimadzu, Columbia, MD) was used for chromatographic separation. An AB SCIEX Qtrap® 5500 mass spectrometer (AB SCIEX, Framingham, MA) operated in negative mode was used for detection. Mass spectral data was processed using MultiQuant® 3.0.1and PeakView® 2.2[1]

In Silico constraint-based modeling, simulation, and atom mapping network generation:

The *E. coli* models used for MFA included iRL2013 [30], iJS2012[28], iDM2014_core (core model derived from iJO1366[37]), and iDM2014 (genome-scale model derived from iJO1366[37]). All constraint-based modeling was conducted in python using COBRApy[41]. Atom mappings were taken from iJS2012[28], the EcoCyc database[32], MetRxn database[31] and KEGG reaction pair database[42]. Metabolic map construction and visualization was done using escher 1.0.0[43].

Metabolic Flux Analysis:

MFA simulations were conducted with MATLAB® and INCA v1.3[44]. A wrapper from INCA to the Cobra Toolbox[45] was written in order to utilize the more precise linear solvers (i.e., glpk http://www.gnu.org/software/glpk/) supported by the Cobra Toolbox. The scripts for the wrapper are available at https://github.com/dmccloskey/genomeScale_MFA_INCA. MIDs were weighted by the standard deviation of biological and analytical replicates (n=6) or the accuracy as determined from unlabeled glucose labeling experiments[46]. 84 MIDs corresponding to the 32 metabolites were included in the fit for the genome-scale model (Supplemental Table S-10). Standard deviations were calculated based off of 95% confidence intervals as described in Antoniewicz 2006[47] as follows: $Standard\ Deviation = \frac{upper\ bound - lower\ bound}{4}$. Observable fluxes were determined as described in Choi 2011[48]. Observable fluxes were those where the estimated flux value was at least four times larger than the 95% confidence interval and did not include the value zero. Standard deviations of observable fluxes were used to compare the precision of each model. Significant difference

between fluxes was determined by the 95% confidence intervals. Further details are described in the supplemental material.

MFA simulations were conducted on a 128 core cluster consisting of dual-socket 2.6 GHz Intel Xeon E5-2670 processors with 64 GB of memory. Flux estimations and confidence interval calculations using iRL2013 and iDM2014 took approximately a minute, and 4 hours, respectively.

Results and Discussion

A tunable MFA Genome-scale model generation workflow:

A genome-scale MFA model was constructed that accounts for metabolic content traditionally omitted from MFA models. This model accounts for cofactor usage, biosynthesis, and salvage pathways, as well as the carbon mappings of biosynthetic precursors, nucleotide phosphate, cofactor, vitamins, and other macromolecules. In addition to the full model, a core model of central carbohydrate metabolism that accounts for the stoichiometry of cofactors, but not the carbon mapping, was also constructed for comparison. The full model, iDM2014, and core model, iDM2014_core, were derived from iJO1366[37] (Figure 1). The core model included the entire central carbohydrate metabolism, amino acid metabolism, and oxidative phosphorylation, as well as necessary transport reactions. Cofactors, vitamins, and other compounds that were contained in the biomass reaction, which could no longer be produced by the model, were replaced with their metabolic precursors (e.g., phosphoenolpyruvate, erythrose-4-phosphate, etc.) to generate a modified biomass reaction composition. Prior to use in MFA simulations, the models were tuned for growth on glucose minimal media. This tuning involved the lumping of linear and/or coupled pathways of peripheral metabolism for the full model into a minimal number of reactions (Figure 1B). “No flux” reactions of peripheral metabolism were removed based on simulation results from FVA[49] and parsimonious network usage (pFBA[50]) for aerobic, glucose minimal media conditions. The iterative procedure of reducing peripheral reactions without changing growth rate predictions is given in Figure 1.

This reduction strategy decreased the number of irreversible network reactions from 3219 in iJO1366 to 679 while retaining the biosynthesis of all components that comprise the biomass reaction composition. It is worthwhile to note that this tuning can and should be performed for any given substrate and media condition of interest to save computational time.

A workflow was developed to integrate carbon mapping transitions of nucleotides, cofactors, vitamins, and other metabolites into the central and peripheral metabolism of the stoichiometric model. Such data sets of experimentally-derived atom transitions that encompass all of *E. coli* metabolism do not yet exist. Some databases of computationally simulated atom transitions that encompass a majority of *E. coli* metabolism are available[31, 32]. However, the atom transitions of core and peripheral metabolism are often incomplete, contain equally probable alternate atom mappings, or biochemically incorrect atom mappings. This makes direct integration of simulated atom mappings with the stoichiometric model problematic without extensive manual curation. To this end, an atom mapping naming convention and procedure was developed that ensured uniqueness of each atom transition in a given reaction and allowed for the building of the atom mappings of peripheral metabolites not accounted for from their precursor metabolites in any database. A schematic of the naming convention and atom mapping procedure for peripheral metabolic reactions is given in Supplemental Figures S-1 and S-2. Other benefits that can be attributed to the procedure generated in this work include the ability to readily combine atom transitions of individual reactions into lumped reactions and the ability to

track the flow of carbon through macromolecule biosynthesis and degradation as precursor units instead of specific carbon positions. The latter is particularly beneficial due to the amount of manual curation of transitions that was often necessary when dealing with incomplete or alternate carbon mappings.

Due to the fact that biochemical references for atom mappings from databases and previously published models are often incomplete and may contain inaccuracies, the robustness of the genome-scale model to incorrect atom mappings and addition or removal of nonessential reactions was explored (supplemental discussion). It was found that the accuracy of the genome-scale model was robust to small errors in atom mappings (i.e., the incorrect assignment of symmetry) and the addition or removal of non-essential reactions (Supplemental Table S7, Supplemental Figure S4 and S5). However, the penalty for including errors in atom mappings and additional reactions was found to be a loss in precision.

A comparison of flux accuracy to previously published MFA models:

The full and core iDM2014 models were compared to previously generated MFA models to evaluate and validate their content (Table 1). MIDs of intracellular intermediates and peripheral metabolites, and measured uptake, secretion, and growth rates from wild-type *E. coli* grown on an 80/20 mixture of 1-^{13}C/U-^{13}C, sampled, and extracted during exponential growth, were used to determine the accuracy (this section) and precision (section 3.3) of MFA flux estimations for all the models described in Table 1. This tracer scheme is commonly used in MFA studies for its ability to successfully resolve a multitude

of paths[25, 26]. There is also evidence demonstrating that the resolution of estimated fluxes when using a genome-scale model is not significantly affected by the tracer scheme used[33]. The models compared include the following: iRL2013 [30], a traditional MFA model of core metabolism; iDM2014_core, a core model that accounts for cofactor mass balances (this work); iJS2012 [28], a previously-published reduced genome-scale model that accounts for cofactor mass balances and all biomass components; and iDM2014, a genome-scale model (as described here) that accounts for cofactor mass balances, their mappings, and all biomass components. The iRL2013 model was specifically chosen as the representative traditional MFA model because it encompasses the content found in the majority of traditional MFA models, and has been extensively validated and utilized [30, 51]. All net fluxes estimated are given in Supplemental Table S-1. The statistics of the fit and flux estimates are given in Tables 2 and 3, respectively.

A comparison of flux estimations for all models examined here were found to be within ranges of previously published flux estimations of wild-type *E. coli* grown on glucose minimal media [2, 25, 28, 30] and similar to results found previously using the same tracer and experimental setup[1]. Only minor net flux differences (i.e., less than 0.005 fold change between intracellular reactions) for observable reactions and reaction identifiers that could be reconciled between models were found (Supplemental Figure S3 and supplemental discussion). Taken together, these findings validate the accuracy of the modeling methods by showing consistency across the models examined.

A comparison of flux precision to previously published MFA models:

The number of observable fluxes (criteria described in materials and methods) and observable flux precision for either all net fluxes or a subset of representative net fluxes from core metabolism was compared for all models (Figure 2 and Table 3). Precision was determined from 95% confidence intervals as described in the methods. A more encompassing biomass reaction and the addition of mass balances for cofactors and additional metabolites did not appear to improve the flux precision. For example, compared to iRL2013, the precision of flux estimates did not improve when using iDM2014_core, which includes cofactor mass balances (0.11 vs. 0.14 for representative net fluxes from core metabolism). Compared to both iRL2013 and iDM2014_core, the precision of flux estimates did not improve when using iJS2012 (0.19 for representative net fluxes from core metabolism), which includes a full biomass objective reaction and cofactor mass balances. However, the addition of mappings for cofactors and additional metabolites did appear to improve flux precision. This is evident when comparing the precision of fluxes estimated using iJS2012, which does not include mappings for cofactors, and iDM2014, which does include mappings for cofactors (i.e., 0.19 vs. 0.10, respectively). This indicates that while increasing the scope of the model can generate a greater number of observable fluxes, the power to resolve those fluxes cannot be realized without corresponding atom mappings for metabolites in peripheral metabolism.

In order to better understand the contribution of including additional MS measurements of cofactors and peripheral metabolites in conjunction with the

carbon mappings for cofactors and peripheral metabolites, the genome-scale model iDM2014 was resimulated without MS measurements for cofactors (results denotes as iDM2014*). First, it is important to note that the chi-squared statistics was not satisfied when using the genome-scale model without the additional MS measurements (Table 3). This may not be true for all use cases of the genome-scale MFA model, but it highlights the importance of having an analytical method capable of measuring the MIDs of peripheral metabolites to compensate for the decreased number of degrees of freedom. Regardless, a reduction in the percentage of observable fluxes (43.2% vs. 42.1%) and a loss of precision was found for observable fluxes (0.22 vs. 1.60) from iDM2014 to iDM2014*. This difference highlights the need for additional measurements in peripheral metabolism to reliably constrain reactions in the periphery. Compared to iJS2012, the percentage of observable fluxes and net flux precision was markedly greater for iDM2014 even without MS measurements of cofactors (35.3% vs. 43.2% and 6.03 vs. 1.6 for the percentage of observable fluxes and average observable flux precision, respectively). This implies that both the MS measurement of cofactors and the carbon mappings themselves contribute to the overall number of observable fluxes and observable flux precision of the genome-scale model iDM2014.

When comparing iDM2014 to iRL2013, a minor improvement in average observable flux precision was found in central carbohydrate reaction fluxes (0.10 vs. 0.11, respectively). When considering all observable fluxes, an improvement in average flux precision was found (0.22 vs. 0.52, respectively). Previous work

has shown that many of the flux ranges for peripheral metabolic pathways are constrained by the full biomass objective[33], which may explain the increase in average observable precision found in this study. However, unlike previous work where intracellular and cofactor MID measurements were not available, this study demonstrates a gain in flux precision for the genome-scale model in core and peripheral metabolism when the analytical capabilities to make such measurements are available (Table 3 and Supplemental Table S-9). Furthermore, the flux ranges of MFA estimated flux values compared to the flux ranges of FVA simulated minimum and maximum values was found to be improved by 2.6% (Supplemental Table S-11). However, it was found that 27.1% of the reactions in peripheral metabolism were significantly different between the two methods. This finding indicates that while the full biomass objective imposes constraints on the reactions of peripheral metabolism, the estimated flux ranges are neither as precise nor accurate unless direct measurement of peripheral MIDs are included. This can be explained by the fact that the percentage of free fluxes in the periphery in iDM2014 was similar to the percentage of free fluxes in central metabolism (38.8% and 40.0%, respectively). Therefore, this study demonstrates that the combination of improved analytical capabilities of MID Max[1] and a condition-specific modeling tuning workflow can allow for precise flux estimations using MFA at a genome-scale.

Increased scope of measured flux values:

iDM2014 captures the recycling of metabolic intermediates from biosynthetic pathways back into central carbohydrate metabolism that are not

captured by traditional MFA models (Figure 4). For example, L-alanine and UDP are recycled back into the network through meurine biosynthetic pathways [52-54]. The amount of meurine recycled per cell division has been estimated to be as high as 30% [54]. Pyruvate can also be recycled from L-serine via serine deaminase. Several salvage pathways such as AMP nucleosidase (which recycles AMP to ribose-5-phosphate and adenine) are identified as active. Glycerol is recycled directly into glycolysis as a byproduct of lipid metabolism. The symmetric metabolite fumarate is a byproduct of several biosynthetic pathways that are reincorporated directly into central carbohydrate metabolism. Fumarate can be recycled from L-aspartate via aspartate transaminase or from (S)-2-[5-Amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamido]succinate via adenylosuccinate lyase. In addition, the genome-scale model captures the multiple biosynthetic routes to synthesize the metabolic intermediate AICAR (Figure 3). AICAR is synthesized from ribose-5-phosphate during histidine and purine biosynthesis along its incorporation into IMP. Similar to previous studies, CO2, glycolate, and formate were also found to be recycled from various biosynthetic processes[33]. Thus, additional sources of variance in the carbon labeling pattern of metabolic intermediates are accounted for when including peripheral metabolism.

A major benefit to using a genome-scale model for MFA is the direct measurement of peripheral metabolic pathway flux (Figure 4, Supplemental Figure 2, Supplemental Table 8). For example, the flux through biosynthetic, interconversion, and degradation pathways of amino acids is measured directly.

Thus, a genome-scale model for MFA could be more directly applicable to analyze the production of amino acids such as L-valine [55-57] and L-threonine [58], which have been targeted for overproduction using *E. coli*. The flux to isoprenoid biosynthesis is directly measured in a genome-scale model for MFA such as iDM2014. Isoprenoids such as carotenoids [59, 60] are high value precursors to the production of a number of flavors, fragrances, and medicines. In particular, the anti-malarial drug artemisinin [61-64] and anti-cancer drug taxol [65, 66] have been targets for overproduction using *E. coli*. In addition, many biosynthetic pathways of cell wall formation are well resolved. The enzymes of these pathways are high value targets for antimicrobial drug design[67-69].

Conclusion:

A workflow for generating genome-scale MFA models and integration with cofactor measurements of MIDs was detailed. The genome-scale model presented here, iDM2014, was tuned for growth on glucose. Other genome-scale models tuned for a desired environmental and/or genetic background can be readily produced using the workflow that was presented. iDM2014 was compared to a core version and previously published MFA models. A 414% gain in the number of calculated fluxes was found. These additional flux measurements include purine, pyrimidine, isoprenoid, methionine, riboflavin, coenzyme A, folate, and other biosynthetic pathways that are excluded from traditional MFA models. Importantly, the gain in the number of calculated fluxes was found without a loss in precision of estimated fluxes of central metabolism. The ability to directly measure an increased scope of reactions without compromising the precision with which those fluxes are estimated should be a valuable resource for the engineering of microorganisms for the synthesis of high value products from renewable energy sources and aid in the discovery and design of novel drug targets.

Acknowledgments

Chapter 7, in full, is a reformatted reprint of “Modeling Method for Increased Precision and Scope of Directly Measurable Fluxes at a Genome-Scale.” McCloskey D, Young JD, Xu S, Palsson BO, Feist AM. Anal Chem. 2016 Apr 5;88(7):3844-52. doi: 10.1021/acs.analchem.5b04914. The dissertation/thesis author was the primary investigator and author of the paper.

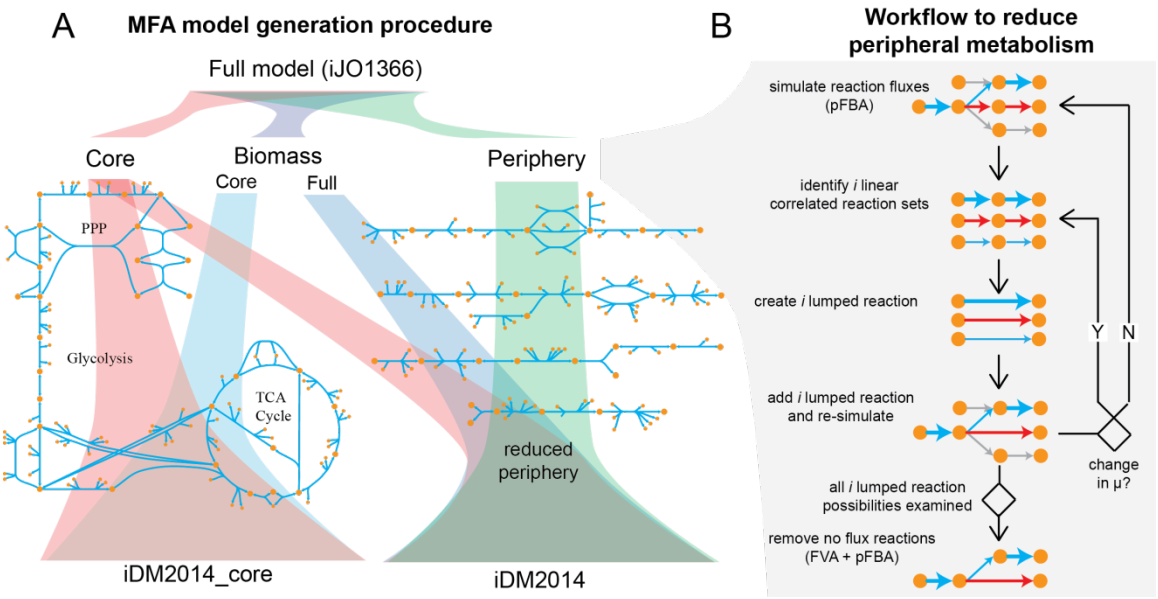

Figure 7.1: Schematic of the genome-scale and core MFA model generation and model generation workflow. (A) Starting from the genome-scale reconstruction iJO1366, core and reduced models of *E. coli* metabolism were generated. The core model (iDM2014_core) included all core metabolic pathways along with a core biomass reaction. The reduced model (iDM2014) included core metabolism, the full biomass reaction, and a reduced peripheral metabolism. (B) A workflow to reduce peripheral metabolism. Peripheral metabolism was iteratively linearized and pruned to maintain an optimal growth rate for the given experimental conditions using constraint-based approaches. Further details of the reduction procedure can be found in the main text and supplemental methods.

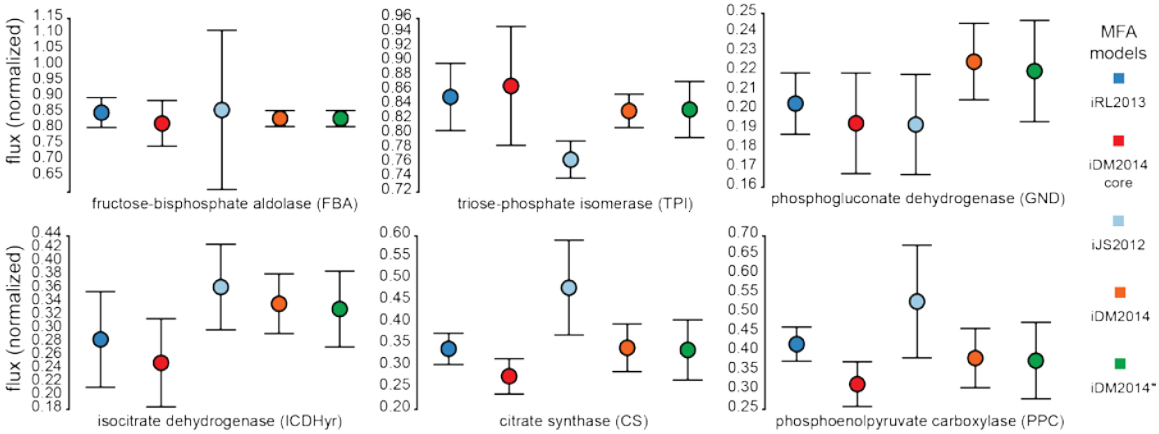

Figure 7.2: Net flux values calculated with different MFA models. Flux predictions, including precision, of key reactions in central carbohydrate metabolism calculated using the same data set and different models. Circles represent the best net flux estimate; whiskers represent +/- the standard deviation as calculated from 95% confidence intervals. All flux values are normalized to net glucose uptake. iDM2014* denotes results when cofactor measurements corresponding to Acetyl-CoA, AMP, ATP, FAD, and UTP were omitted from the flux estimation using iDM2014.

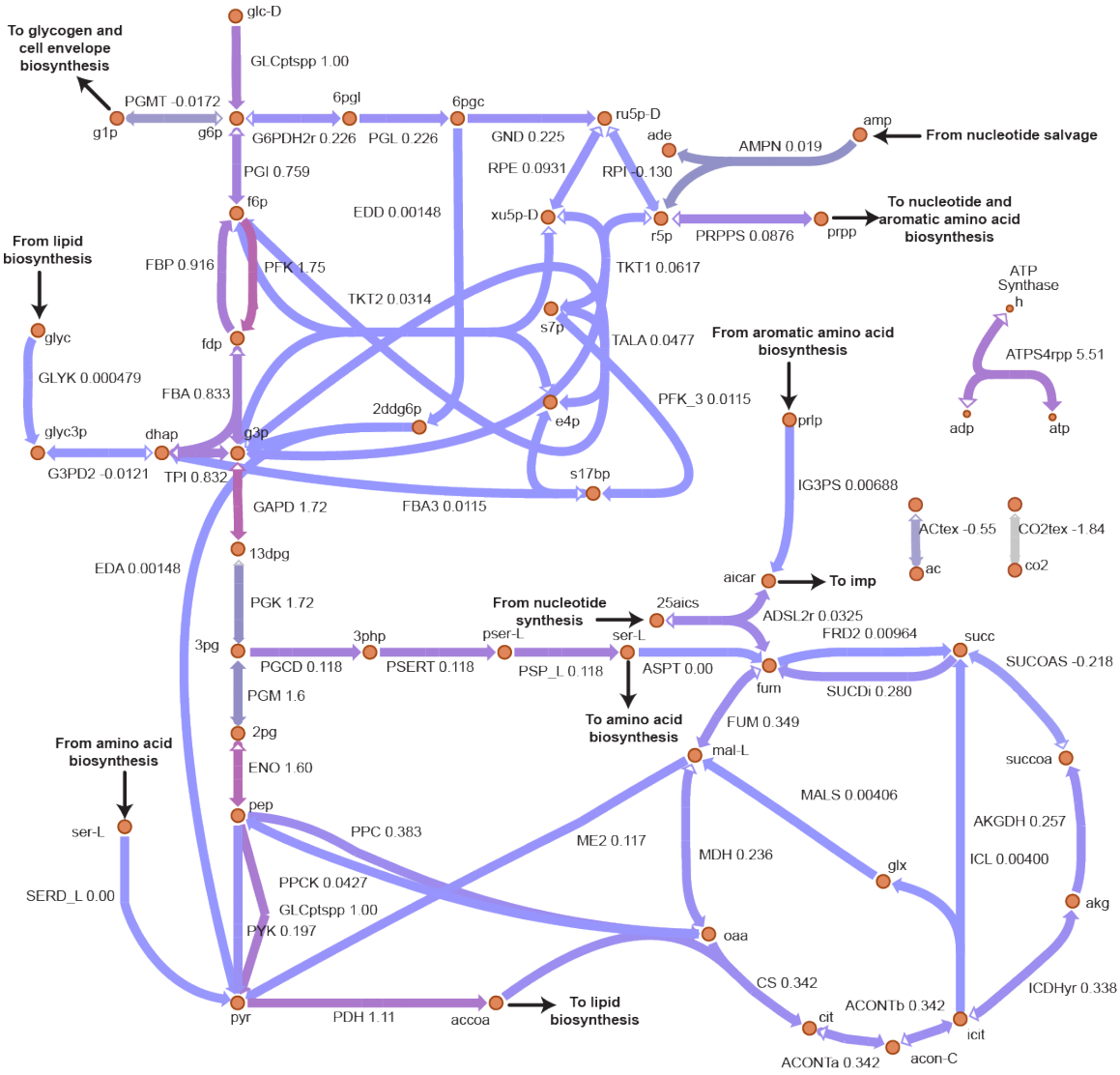

Figure 7.3: A predicted flux map for central carbohydrate metabolism for wild-type *E. coli* using the iDM2014 MFA model. The flux map was generated using measured data for *E. coli* growing aerobically in glucose M9 minimal media. Shown are core metabolic reactions along with the entry points of several recycled intermediate metabolites for biosynthetic and salvage pathways of peripheral metabolism. Also shown are the multiple biosynthetic routes to synthesize AICAR and glycine that are not captured by traditional MFA models. A complete list of reaction and metabolite abbreviations is given in S-5.

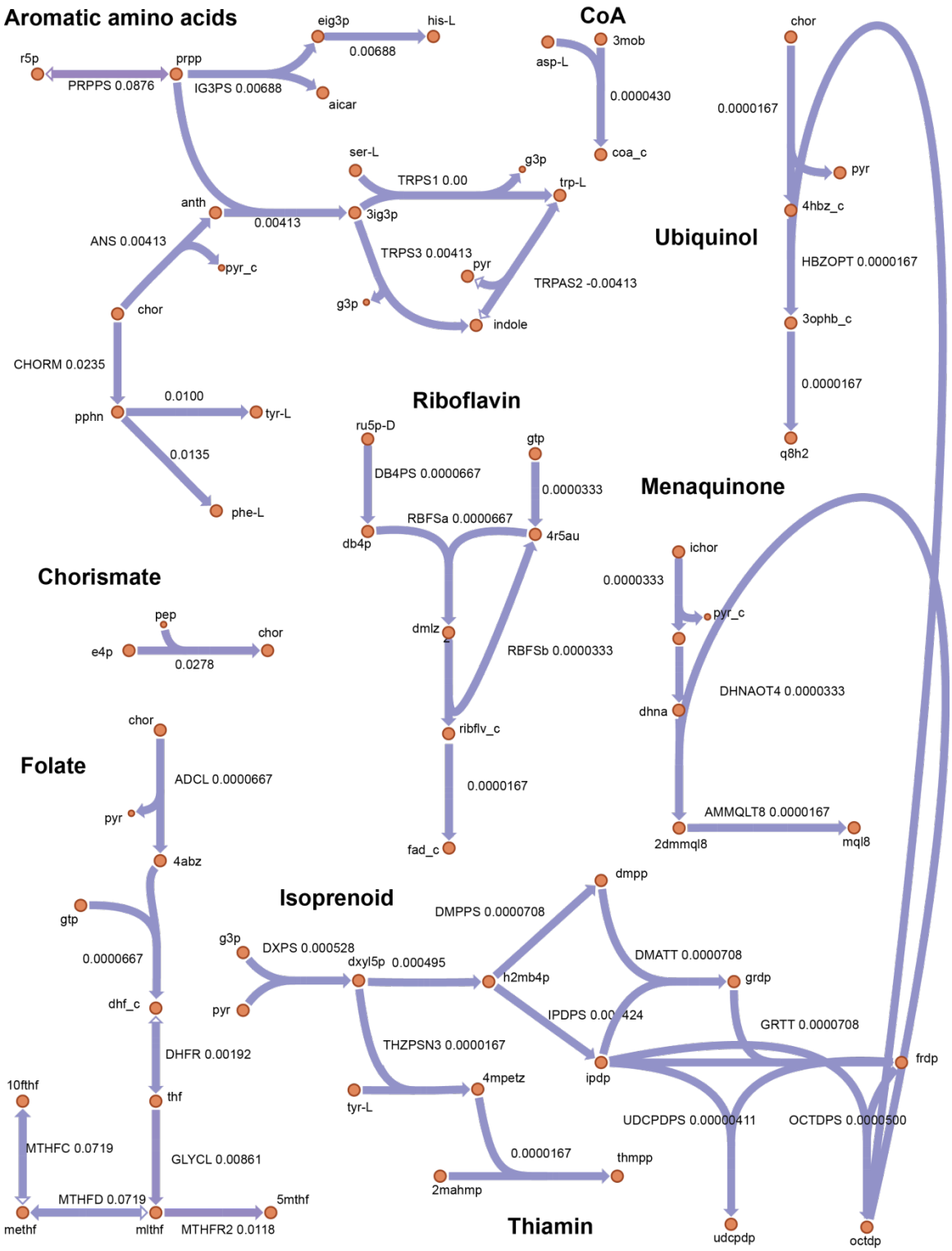

Figure 7.4: A predicted flux map for select pathways in peripheral metabolism for wild-type *E. coli* using the iDM2014 MFA model that have been targets of metabolic engineering. The flux map was generated using measured data for *E. coli* growing aerobically in glucose M9 minimal media. The name for lumped reactions has been omitted. A complete list of reaction and metabolite abbreviations is given in S-5.

Table 7.1: MFA Model statistics

Model	irreversible reactions	net reactions	metabolites	reactions with C-mappings	metabolites with C-mappings	full biomass reaction	cofactor balance	cofactor mappings	reference
iRL2013	97	75	74	89	62	No	No	No	[30]
iDM2014_core	430	260	262	280	157	No	Yes	No	This study
iJS2012	353	266	190	265	127	Yes	Yes	No	[28]
iDM2014	679	537	429	564	536	Yes	Yes	Yes	This study

Table 7.2: Flux estimation statistics. The degrees of freedom (DOF) are calculated as follows: number of measured fluxes + number of measured fragment isotopes – number of free fluxes. iDM2014* denotes results when cofactor measurements corresponding to Acetyl-CoA, AMP, ATP, FAD, and UMP were omitted from the flux estimation using iDM2014.

model	**fluxes included in the estimation**	**fragments included in the estimation**	**χ^2**	**χ^2 accepted**	**DOF**
iRL2013	3	71	280.3	True	283
iDM2014_core	3	72	172.9	True	163
iJS2012	3	72	155.6	True	161
iDM2014	3	86	178.7	True	195
iDM2014*	3	72	129.4	False	60

Table 7.3: Comparison of estimated fluxes between models. Representative fluxes are given in the supplemental information. iDM2014* denotes results when cofactor measurements corresponding to Acetyl-CoA, AMP, ATP, FAD, and UMP were omitted from the flux estimation using iDM2014.

	all net fluxes		**representative core net fluxes (n=34)**	
Model	**Number of observable fluxes (% of net fluxes)**	**Average observable precision)**	**number of observable fluxes (% of net fluxes)**	**Average observable precision**
iRL2013	56 (74.7%)	0.52	19 (55.9%)	0.11
iDM2014_core	91 (35.0%)	3.42	19 (55.9%)	0.14
iJS2012	94 (35.3%)	6.03	14 (41.2%)	0.19
iDM2014	232 (43.2%)	0.22	17 (50.0%)	0.10
iDM2014*	226 (42.1%)	1.6	16 (47.1%)	0.11

References:

1. McCloskey, D., Young, J.D., Xu, S., Palsson, B.Ø. & Feist, A.M. MID Max: A LC-MS/MS method for measuring the precursor and product mass isotopomer distributions (MIDs) of metabolic intermediates and cofactors for metabolic flux analysis (MFA) applications. *Analytical Chemistry* (2015).

2. Toya, Y. 13C-metabolic flux analysis for batch culture of Escherichia coli and its Pyk and Pgi gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. *Biotechnol Prog* **26**, 975-992 (2010).

3. Fong, S.S., Nanchen, A., Palsson, B.O. & Sauer, U. Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. *J Biol Chem* **281**, 8024-8033 (2006).

4. Iwatani, S. Determination of metabolic flux changes during fed-batch cultivation from measurements of intracellular amino acids by LC-MS/MS. *J Biotechnol* **128**, 93-111 (2007).

5. Chen, X., Alonso, A.P., Allen, D.K., Reed, J.L. & Shachar-Hill, Y. Synergy between (13)C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in E. coli. *Metab Eng* **13**, 38-48 (2011).

6. Ewald, J.C., Matt, T. & Zamboni, N. The integrated response of primary metabolites to gene deletions and the environment. *Mol Biosyst* **9**, 440-446 (2013).

7. Peng, L., Arauzo-Bravo, M.J. & Shimizu, K. Metabolic flux analysis for a ppc mutant Escherichia coli based on 13C-labelling experiments together with enzyme activity assays and intracellular metabolite measurements. *FEMS Microbiol Lett* **235**, 17-23 (2004).

8. Swarup, A., Lu, J., DeWoody, K.C. & Antoniewicz, M.R. Metabolic network reconstruction, growth characterization and 13C-metabolic flux analysis of the extremophile Thermus thermophilus HB8. *Metabolic Engineering* **24**, 173-180 (2014).

9. Au, J., Choi, J., Jones, S.W., Venkataramanan, K.P. & Antoniewicz, M.R. Parallel labeling experiments validate Clostridium acetobutylicum metabolic network model for 13C metabolic flux analysis. *Metabolic Engineering* **26**, 23-33 (2014).

10. Yim, H. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* **7**, 445-452 (2011).

11. Antoniewicz, M.R. Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol. *Metabolic Engineering* **9**, 277-292 (2007).

12. Bartek, T. Comparative 13C Metabolic Flux Analysis of Pyruvate Dehydrogenase Complex-Deficient, l-Valine-Producing Corynebacterium glutamicum. *Applied and Environmental Microbiology* **77**, 6644-6652 (2011).

13. Shirai, T. Study on roles of anaplerotic pathways in glutamate overproduction of Corynebacterium glutamicum by metabolic flux analysis. *Microbial Cell Factories* **6**, 19 (2007).

14. Umakoshi, M. Improving protein secretion of a transglutaminase-secreting Corynebacterium glutamicum recombinant strain on the basis of 13C metabolic flux analysis. *Journal of Bioscience and Bioengineering* **112**, 595-601 (2011).

15. Jorda, J. Metabolic flux profiling of recombinant protein secreting Pichia pastoris growing on glucose:methanol mixtures. *Microbial Cell Factories* **11**, 57 (2012).

16. Lu, S., Eiteman, M.A. & Altman, E. Effect of CO2 on succinate production in dual-phase Escherichia coli fermentations. *Journal of Biotechnology* **143**, 213-223 (2009).

17. Templeton, N. The impact of anti-apoptotic gene Bcl-2Δ expression on CHO central metabolism. *Metabolic Engineering* **25**, 92-102 (2014).

18. McCloskey, D., Palsson, B.O. & Feist, A.M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* **9**, 661 (2013).

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

20. Thiele, I. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. *PLoS ONE* **7**, e45635 (2012).

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

22. Liu, J. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Systems Biology* **8**, 110 (2014).

23. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol* **3**, 119 (2007).

24. Zamboni, N., Fendt, S.M., Ruhl, M. & Sauer, U. (13)C-based metabolic flux analysis. *Nat Protoc* **4**, 878-892 (2009).

25. Fischer, E., Zamboni, N. & Sauer, U. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Anal Biochem* **325**, 308-316 (2004).

26. Fischer, E. & Sauer, U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nat Genet* **37**, 636-640 (2005).

27. Zamboni, N., Fischer, E. & Sauer, U. FiatFlux--a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* **6**, 209 (2005).

28. Schellenberger, J. Predicting outcomes of steady-state (1)(3)C isotope tracing experiments using Monte Carlo sampling. *BMC Syst Biol* **6**, 9 (2012).

29. Toya, Y. Direct measurement of isotopomer of intracellular metabolites using capillary electrophoresis time-of-flight mass spectrometry for efficient metabolic flux analysis. *J Chromatogr A* **1159**, 134-141 (2007).

30. Leighty, R.W. & Antoniewicz, M.R. COMPLETE-MFA: complementary parallel labeling experiments technique for metabolic flux analysis. *Metab Eng* **20**, 49-55 (2013).

31. Ravikirthi, P., Suthers, P.F. & Maranas, C.D. Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and Bioengineering* **108**, 1372-1382 (2011).

32. Latendresse, M., Malerich, J.P., Travers, M. & Karp, P.D. Accurate Atom-Mapping Computation for Biochemical Reactions. *Journal of Chemical Information and Modeling* **52**, 2970-2982 (2012).

33. Gopalakrishnan, S. & Maranas, C.D. 13C metabolic flux analysis at a genome-scale. *Metabolic Engineering* **32**, 12-22 (2015).

34. Kuepfer, L., Sauer, U. & Blank, L.M. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res* **15**, 1421-1430 (2005).

35. Choi, H.S., Kim, T.Y., Lee, D.-Y. & Lee, S.Y. Incorporating metabolic flux ratios into constraint-based flux analysis by using artificial metabolites and converging ratio determinants. *Journal of Biotechnology* **129**, 696-705 (2007).

36. García Martín, H. A Method to Constrain Genome-Scale Models with ¹³C Labeling Data. *PLoS Comput Biol* **11**, e1004363 (2015).

37. Orth, J.D. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* **7**, 535 (2011).

38. Sambrook, J., and D. W. Russell Molecular cloning: a laboratory manual, 3rd ed., vol. A2.2. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001).

39. Fong, S.S. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnology and Bioengineering* **91**, 643-648 (2005).

40. McCloskey, D., Utrilla, J., Naviaux, R., Palsson, B. & Feist, A. Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. *Metabolomics*, 1-12 (2014).

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

42. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

43. King ZA, D.A., Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO. Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLOS Computational Biology* **Accepted** (2015).

44. Young, J.D. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* **30**, 1333-1335 (2014).

45. Schellenberger, J. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290-1307 (2011).

46. Young, J.D., Allen, D.K. & Morgan, J.A. Isotopomer measurement techniques in metabolic flux analysis II: mass spectrometry. *Methods Mol Biol* **1083**, 85-108 (2014).

47. Antoniewicz, M.R., Kelleher, J.K. & Stephanopoulos, G. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metab Eng* **8**, 324-337 (2006).

48. Choi, J. & Antoniewicz, M.R. Tandem mass spectrometry: a novel approach for metabolic flux analysis. *Metab Eng* **13**, 225-233 (2011).

49. Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**, 264-276 (2003).

50. Lewis, N.E. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* **6**, 390 (2010).

51. Crown, S.B., Long, C.P. & Antoniewicz, M.R. Integrated 13C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. *Metab Eng* **28**, 151-158 (2015).

52. Goodell, E.W. Recycling of murein by Escherichia coli. *Journal of Bacteriology* **163**, 305-310 (1985).

53. Goodell, E.W. & Schwarz, U. Release of cell wall peptides into culture medium by exponentially growing Escherichia coli. *Journal of Bacteriology* **162**, 391-397 (1985).

54. Jacobs, C., Huang, L.J., Bartowsky, E., Normark, S. & Park, J.T. Bacterial cell wall recycling provides cytosolic muropeptides as effectors for beta-lactamase induction. *The EMBO Journal* **13**, 4684-4694 (1994).

55. Park, J.H., Lee, K.H., Kim, T.Y. & Lee, S.Y. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci U S A* **104**, 7797-7802 (2007).

56. Park, J.H., Kim, T.Y., Lee, K.H. & Lee, S.Y. Fed-batch culture of Escherichia coli for L-valine production based on in silico flux response analysis. *Biotechnol Bioeng* **108**, 934-946 (2011).

57. Park, J.H., Jang, Y.S., Lee, J.W. & Lee, S.Y. Escherichia coli W as a new platform strain for the enhanced production of L-valine by systems metabolic engineering. *Biotechnol Bioeng* **108**, 1140-1147 (2011).

58. Lee, K.H., Park, J.H., Kim, T.Y., Kim, H.U. & Lee, S.Y. Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol* **3**, 149 (2007).

59. Matthews, P.D. & Wurtzel, E.T. Metabolic engineering of carotenoid accumulation in Escherichia coli by modulation of the isoprenoid precursor pool with expression of deoxyxylulose phosphate synthase. *Appl Microbiol Biotechnol* **53**, 396-400 (2000).

60. Yuan, L.Z., Rouvière, P.E., LaRossa, R.A. & Suh, W. Chromosomal promoter replacement of the isoprenoid pathway for enhancing carotenoid production in E. coli. *Metabolic Engineering* **8**, 79-90 (2006).

61. Anthony, J.R. Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Escherichia coli for production of the anti-malarial drug precursor amorpha-4,11-diene. *Metab Eng* **11**, 13-19 (2009).

62. Martin, V.J., Pitera, D.J., Withers, S.T., Newman, J.D. & Keasling, J.D. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat Biotechnol* **21**, 796-802 (2003).

63. Newman, J.D. High-level production of amorpha-4,11-diene in a two-phase partitioning bioreactor of metabolically engineered Escherichia coli. *Biotechnol Bioeng* **95**, 684-691 (2006).

64. Pitera, D.J., Paddon, C.J., Newman, J.D. & Keasling, J.D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab Eng* **9**, 193-207 (2007).

65. Ajikumar, P.K. Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli. *Science* **330**, 70-74 (2010).

66. Huang, Q., Roessner, C.A., Croteau, R. & Scott, A.I. Engineering Escherichia coli for the synthesis of taxadiene, a key intermediate in the biosynthesis of taxol. *Bioorganic & Medicinal Chemistry* **9**, 2237-2242 (2001).

67. Couce, A. Genomewide Overexpression Screen for Fosfomycin Resistance in Escherichia coli: MurA Confers Clinical Resistance at Low Fitness Cost. *Antimicrobial Agents and Chemotherapy* **56**, 2767-2769 (2012).

68. Kim, D.H. Characterization of a Cys115 to Asp Substitution in the Escherichia coli Cell Wall Biosynthetic Enzyme UDP-GlcNAc Enolpyruvyl Transferase (MurA) That Confers Resistance to Inactivation by the Antibiotic Fosfomycin. *Biochemistry* **35**, 4923-4928 (1996).

69. Schönbrunn, E., Eschenburg, S., Krekel, F., Luger, K. & Amrhein, N. Role of the Loop Containing Residue 115 in the Induced-Fit Mechanism of the Bacterial Cell Wall Biosynthetic Enzyme MurA‡. *Biochemistry* **39**, 2164-2173 (2000).

CHAPTER 8:

Laboratory Evolution of Gene Knockout Strains Reveals Fundamental Principles of Adaptation

Abstract:

Adaptive laboratory evolution (ALE) enables a new line of biological inquiry. ALE was used to study re-optimization of growth fitness of a pre-evolved Escherichia coli K-12 MG1655 in response to knock-out (KO) of major metabolic genes. Metabolomic, fluxomic, transcriptomic, and genomic resequencing data allowed for detailed analysis of changes between the reference (Ref), unevolved KO (uKO), and evolved KO (eKO) strains. Changes in these data sets revealed fundamental principles underlying adaptation. First, the majority of cellular components in the eKO strains reverted to levels near those in the Ref strain, representing drivers towards optimality. Second, the remaining cellular components took on variable levels in replicate endpoints, representing alternate optimal states. Third, KOs imbalanced metabolite concentrations that affected transcription factor activity and thus gene expression. Mutations selected during ALE reprogrammed transcription factor responses in the Ref strain that malfunctioned after KO to produce the new optimal homeostatic state.

Introduction:

Whole genome sequences and their functional annotations have now existed for over 20 years (Fleischmann et al., 1995). The functions of many of the gene products in model organisms have been isolated and experimentally characterized. In contrast to our understanding of individual cellular components, our knowledge of condition-specific functions and biomolecular interactions that comprise biochemical networks is still limited. Systems biology has arisen as a field to address these challenges in a bottom-up mechanistic fashion with a suite of computational capabilities to assess network functions (Bordbar et al., 2014; O'Brien et al., 2015).

Even with advances in experimental laboratory methods and computer models of network functions, many gaps in our knowledge about biochemistry remain. A method for discovering novel gene functions and novel biomolecular interactions involves the use of adaptive laboratory evolution (ALE). ALE is an experimental method that introduces a selection pressure (e.g., growth rate selection) in a controlled environmental setting (Dragosits and Mattanovich, 2013; Plucain et al., 2014; Tenaillon et al., 2016). Using ALE, organisms can be perturbed from their evolutionary optimized homeostatic states, and their re-adjustments can be studied during the course of adaptation to reveal novel and non intuitive component functions and interactions. ALE has been applied to study adaptation to new environments (Applebee et al., 2011; Fong et al., 2005a; Ibarra et al., 2002; Tenaillon et al., 2012) and adaptation to genetic manipulations (Charusanti et al., 2010; Fong et al., 2006). These studies have yielded

fundamental biological insight that is difficult to gain without direct observation of the evolution process (Conrad et al., 2011; Pál et al., 2014). Automation of ALE has increased its scale and accuracy (LaCroix et al., 2015).

In this study, the adaptation of *E. coli* K-12 MG1655 to major metabolic perturbations was examined. First, a novel experimental design is introduced (Figure 1). Major metabolic functions were removed (i.e., perturbations) through a series of gene KOs in a pre-evolved *E. coli* strain (i.e., optimized system). Automated ALE was then used to evolve the organism from the perturbed state to a new optimized states (i.e., recovery). Multi-omic data sets were generated (i.e., system components) in the optimized, perturbed, and recovered states. Second, a high level analysis of the data sets is performed (Figure 2). Multivariate statistics were used to decompose omics data sets into dominant 'modes' (Figure 2A). Two primary modes emerged: 1) the transition between initial sub-optimal and re-optimized system of the majority of systems components, and 2) alternate re-optimized system configurations, respectively. Third, a model of systems adaptation is developed. Based on detailed analyses and case studies (Figures 3-6, Figures S2-7, and Tables S2-7), an overall model of adaptation to metabolic perturbation is constructed (Figure 7A). In this model, imbalances in metabolite levels, resulting from altered metabolic fluxes, triggered a multitude of network regulatory responses that were readjusted by mutations selected for during adaptive evolution (subsections i-vii). Fourth, principles underlying adaptation to metabolic perturbations are revealed (Figure 7B). The Results section is organized based on these four steps.

Materials and Methods:

Experimental Model and Subject Details:

A glucose, 37°C, evolved *E.coli* derived from *E. coli* K-12 MG1655 (ATCC 700926)(LaCroix et al., 2015; Sandberg et al., 2014) served as the starting strain. Lambda-red mediated DNA mutagenesis (Datsenko and Wanner, 2000) was used to create the knockout strains (DNA mutagenesis and PCR confirmation primers are given in Table S2). Knockouts were confirmed by PCR and DNA resequencing. Genes gnd, ptsH, ptsI, crr, sdhC, sdhA, sdhD, sdhC, tpiA, and pgi encoding for the reactions of 6-phosphogluconate dehydrogenase (GND), phosphotransferase sugar import (GLCptspp), succinate dehydrogenase complex (SUCDi), triophosphate isomerase (TPI), and phosphoglucose isomerase (PGI) were removed. PPC was also deleted, but resulted in an auxotrophy for asp-L, and was not included in the study. Genes aceE, aceF, zwf, and atpI-A encoding for the reactions of PDH, G6PDH2r, and ATPS4rpp could not be removed using the method of Datsenko, et. al. All cultures were grown in 25 mL of unlabeled or labeled glucose M9 minimal media (Sambrook and Russell, 2001) with trace elements (Fong et al., 2005b) and sampled from a heat block in 50 mL autoclaved tubes that were maintained at 37°C and aerated using magnetics.

Method Details:

<u>Biological material, analytical reagents, and experimental conditions</u>

Materials and Reagents:

Uniformly labeled ^{13}C glucose and 1-^{13}C glucose was purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA). Unlabeled glucose and other media components were purchased from Sigma-Aldrich (St. Louis, MO). LC-MS reagents were purchased from Honeywell Burdick & Jackson® (Muskegon, MI), Fisher Scientific (Pittsburgh, PA) and Sigma-Aldrich (St. Louis, MO).

Reaction knockout selection:

iJO1366 (Orth et al., 2011) was used as the metabolic model for *E. coli* metabolism; GLPK (version 4.57) was used as the linear program solver. MCMC sampling (Schellenberger and Palsson, 2009) was used to predict the flux distribution of the optimized reference strain. Uptake, secretion, and growth rates were constrained to the measured average value ± SD. Potential reaction deletions were ranked by 1) averaged sampled flux, 2) the number of immediate upstream and downstream metabolites that could be measured, 3) the number of genes required to produce a functional enzyme. Reactions involved in sampling loops, that were spontaneous, were computationally or experimentally essential, or were not actively expressed under the experimental growth conditions were not included in the analysis. Also, reactions that would require more than one genetic alteration to abolish activity were excluded. The top 9 reactions deletions from the rank ordered set of reactions that met the above criteria were chosen for implementation.

Adaptive laboratory evolution (ALE):

Cultures were serially propagated (100 µL passage volume) in 15 mL (working volume) flasks of M9 minimal medium with 4 g/L glucose, kept at 37°C and well-mixed for full aeration. An automated system passed the cultures to fresh flasks once they had reached an OD_{600} of 0.3 (Tecan Sunrise plate reader, equivalent to an OD_{600} of ~1 on a traditional spectrophotometer with a 1 cm path length), a point at which nutrients were still in excess and exponential growth had not started to taper off (confirmed with growth curves and HPLC measurements). Four OD_{600} measurements were taken from each flask, and the slope of $\ln(OD_{600})$ vs. time determined the culture growth rates. A cubic interpolating spline constrained to be monotonically increasing was fit to these growth rates to obtain the fitness trajectory curves.

Multi-omics data processing

Phenomics:

Physiological measurements for culture density were measured at 600 nm absorbance with a spectrophotometer and correlated to cell biomass. Samples to determine substrate uptake and secretion were filtered through a 0.22 µm filter(PVDF, Millipore) and measured using refractive index (RI) detection by HPLC (Agilent 12600 Infinity) with a Bio-Rad Aminex HPX87-H ion exclusion column (injection volume, 10 ul) and 5 mM H2SO4 as the mobile phase (0.5 ml/min, 45°C). Growth, uptake, and secretion rates were calculated from a minimum of four steady-state time-points.

LC-MS/MS instrumentation and data processing:

Metabolites were acquired and quantified on an AB SCIEX Qtrap® 5500 mass spectrometer (AB SCIEX, Framingham, MA) and processed using MultiQuant® 3.0.1 as described previously(McCloskey et al., 2015). Mass isotopomer distributions (MIDs) were acquired on the same instrument and processed using MultiQuant® 3.0.1 and PeakView® 2.2 as described previously (McCloskey et al., 2016a).

Metabolomics:

Internal standards were generated as described previously (McCloskey et al., 2014a). All samples and calibrators were spiked with the same amount of internal standard taken from the same batch of internal standards. Calibration curves were ran before and after all biological and analytical replicates. The consistency of quantification between calibration curves was checked by running a Quality Control sample that was composed of all biological replicates twice a day. Solvent blanks were injected every ninth sample to check for carryover. System suitability tests were injected daily to check instrument performance.

Metabolomics samples were acquired from triplicate cultures (1 mL of cell broth at an OD600 ~ 1.0) using a previously described method(McCloskey et al., 2014b). A pooled sample of the filtered medium that was re-sampled using the FSF filtration technique and processed in the same way as the biological triplicates was used as an analytical blank. Extracts obtained from triplicate cultures and re-filtered medium were analyzed in duplicate. The intracellular values reported, unless otherwise noted, are derived from the average of the biological triplicates (n=6). Metabolites in the pooled filtered medium with a

concentration greater than 80% of that found in the triplicate samples were not analyzed. In addition, metabolites that were found to have a quantifiable variability (RSD >= 50%) in the Quality Control samples or any individual components with an RSD >= 80 were not used for analysis.

Missing values were imputed using a bootstrapping approach as coded in the R package Amelia II(Honaker et al., 2011) (version 1.7.4, 1000 imputations). Remaining missing values were approximated as ½ the lower limit of quantification for the metabolite normalized to the biomass of the sample. Prior to statistical analyses, metabolite concentrations were log normalized to generate an approximately normal distribution using the R package LMGene(Rocke et al.) (version 3.3, “mult”=”TRUE”, “lowessnorm”=”FALSE”). A Bonferroni-adjusted p-value cutoff of 0.01 as calculated from a Student’s t-test was used to determine significance between metabolite concentration levels. The glog-normalized values or the median-normalized values to the reference strain (FC-median vs. ref) were used for downstream statistical analyses.

Fluxomics:

Fluxomics samples were acquired from triplicate cultures (10 mL of cell broth at an OD600 ~ 1.0) using a modified version of the FSF technique as described previously(McCloskey et al., 2016a). MIDs were calculated from biological triplicates ran in analytical duplicate (n=6). MIDs with an RSD greater than 50 were excluded. In addition, MIDs with a mass that was found to have a signal greater than 80% in unlabeled or blank samples were excluded. A previously validated genome-scale MFA model of *E. coli* with minimal alterations

was used for all MFA estimations using INCA(Young, 2014) (version 1.4) as described previously(McCloskey et al., 2016b). The model was constrained using MIDs as well as measured growth, uptake, and secretion rates. Best flux values that were used to calculate the 95% confidence intervals were estimated from 500 restarts.

The 95% confidences intervals were used as lower and upper bound reaction constraints for further constraint-based analyses. MFA derived constraints that violated optimality were discarded and resampled. The descriptive statistics (i.e., mean, median, interquartile ranges, min, max, etc.) for each reaction for each model were calculated from 5000 points sampled from 5000 steps using optGpSampler(Megchelenbrink et al., 2014)(version 1.1), which resulted in an approximate mixed fraction of 0.5 for all models. A permuted pvalue < 0.05 and geometric fold-change of sampled flux values > 0.001 were used to determine differential flux levels, differential metabolite utilization levels, and differential subsystem utilization levels between models. Demand reactions and reactions corresponding to Unassigned, Transport; Outer Membrane Porin, Transport; Inner Membrane, Inorganic Ion Transport and Metabolism, Transport; Outer Membrane, Nucleotide Salvage Pathway, Oxidative Phosphorylation were excluded from differential flux analysis. The geometric fold-change of the mean between models and the reference model were used for hierarchical clustering; the median, interquartile ranges, min, and max values of each sampling distribution for each reaction and model were used as representative samples for downstream statistical analyses.

Transcriptomics:

Total RNA was sampled from triplicate cultures (3 mL of cell broth at an OD600 ~ 1.0) and immediately added to 2 volumes Qiagen RNA-protect Bacteria Reagent (6 mL), vortexed for 5 seconds, incubated at room temperature for 5 min, and immediately centrifuged for 10 min at 17,500 RPMs. The supernatant was decanted and the cell pellet was stored in the -80°C. Cell pellets were thawed and incubated with Readylyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20 minutes at 37°C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30 minutes at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA-nano chip on a bioanalyzer. The rRNA was removed using Epicentre's Ribo-Zero rRNA removal kit for Gram Negative Bacteria. a KAPA Stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used following the manufacturer's protocol to create sequencing libraries with an average insert length of around ~300 bp for two of the three biological replicates. Libraries were ran on a MiSeq and/or HiSeq (illumina).

RNA-Seq reads were aligned using Bowtie(Langmead et al., 2009) (version 1.1.2 with default parameters). Expression levels for individual samples were quantified using Cufflinks(Trapnell et al., 2010)(version 2.2.1, library type fr-firststrand) Quality of the reads was assessed by tracking the percentage of unmapped reads and expression level of genes that mapped to the ribosomal gene loci rrsA-F and rrlA-F. All samples had a percentage of unmapped reads

less than 7%. Differential expression levels for each condition (n=2 per condition) compared to either the starting strain or initial knockout strain were calculated using Cuffdiff(Trapnell et al., 2010)(version 2.2.1, library type fr-firststrand, library norm geometric). Genes with an 0.05 FDR-adjusted p-value less than 0.01 were considered differentially expressed. Expression levels for individual samples for all combinations of conditions tested in down-stream statistical analyses were normalized using Cuffnorm(Trapnell et al., 2010)(version 2.2.1, library type fr-firststrand, library norm geometric). Genes with unmapped reads were imputed using a bootstrapping approach as coded in the R package Amelia II (version 1.7.4, 1000 imputations). Remaining missing values were filled using the minimum expression level of the data set. Normalized FPKM values for gene expression were log2 normalized to generate an approximately normal distribution prior to any statistical analysis. All replicates for a given condition were found to have a pair-wise Pearson correlation coefficient of 0.95 or greater.

DNA resequencing:

Total DNA was sample from an overnight culture (1 mL of cell broth at an OD600 of ~2.0) and immediately centrifuged for 5 min at 8000 RPMs. The supernatant was decanted and the cell pellet was frozen in the -80C. Genomic DNA was isolated using a Nucleospin Tissue kit (Macherey Nagel 740952.50) following the manufacturer's protocol, including treatment with RNase A. Resequencing libraries were prepared using a Nextera XT kit (Illumina FC-131-

1024) following the manufacturer's protocol. Libraries were ran on a MiSeq (illumina).

DNA resequencing reads were aligned to the *E. coli* reference genome (U00096.2, genbank) using Breseq (Deatherage and Barrick, 2014)(version 0.26.0) as populations. Mutations with a frequency of less than 0.1, p-value greater than 0.01, or quality score less than 6.0 were removed from the analysis. In addition, genes corresponding to crl, insertion elements (i.e, insH1, insB1, and insA), and the rhs and rsx gene loci were not considered for analysis due to repetitive regions that appear to cause frequent miscalls when using Breseq. mRNA and peptide sequence changes were predicted using BioPython (https://github.com/biopython/biopython.github.io/). Large regions of DNA (minimum of 200 consecutive indices) where the coverage was two times greater than the average coverage of the sample were considered duplications.

Quantification and Statistical Analysis:

Individual -omics statistical, graph, and modeling data analyses

Structural analysis:

Corresponding PDB files for genes with a mutation of interested were downloaded from PDB (Berman et al., 2003, 2000). Structural models for genes for which there were no corresponding PDB files were taken from I-TASSER generated homology models (Xu and Zhang, 2013) or generated using the I-TASSER protocol (Wu et al., 2007). The BioPython predicted sequence changes and important protein features as listed in EcoCyc (Keseler et al., 2013) were visualized and annotated using VMD (Humphrey et al., 1996).

System component statistical feature identification analyses:

Network components (i.e., RNAseq, metabolomics, fluxomics, genomics) were pre-processed as described above, and subjected to a feature identification analysis pipeline. Network components for each lineage were first subjected to a differential test (ref vs. KO, KO vs. endpoints, ref vs. endpoints, and endpoints vs. endpoints). The criteria for significance for each of the data types are detailed below. Metabolomics: pvalue < 0.01 and 0.5<fold_change<2.0 as calculated from a t-test of the g-log normalized metabolite concentrations. Transcriptomics: qvalue (0.05 FDR corrected pvalue) and abs (log2(fold change))>1.0 as calculated by Cuffdiff. Fluxomics: pvalue < 0.01 and abs (geometric fold_change)>0.001 as calculated from re-sampled flux distributions that were constrained by the 95% confidence intervals derived from estimated MFA flux bounds (demand reactions and reactions in subsystems corresponding to Unassigned, Transport; Outer Membrane Porin, Transport; Inner Membrane, Inorganic Ion Transport and Metabolism, Transport; Outer Membrane, Nucleotide Salvage Pathway, Oxidative Phosphorylation were excluded). Mutations: frequency > 0.1 (mutations in the reference strain and in repetitive regions were excluded). Components that met the significance criteria for any combination of comparisons from the differential test were used in the pairwise PLS-DA analyses and profile matching. Counts of significant components for each lineage were based on components that met the significance criteria for Ref vs. eRef, or uKO vs. eKO.

Network components for each lineage were subjected to pairwise PLS-DA analyses (ref vs. KO, KO vs. endpoints, ref vs. endpoints, and endpoints vs. endpoints). The components with a loadings 1 magnitude within the top 25% of all components and correlation coefficient > 0.88 for different combinations of comparison were selected from the pair-wise PLS-DA analyses.

Network components for each lineage were subjected to profile matching. System component levels between Ref, eKO, and uKO were correlated (Pearson's R) to six profiles in both positive and negative directions. *novel -, novel +, overcompensation -, overcompensation +, partially-restored -, partially-restored +, reinforced -, reinforced +, restored -, restored +, unrestored -, unrestored* + profiles were encoded in integer form as 1-1-0, 0-0-1, 1-0-2, 1-2-0, 2-0-1, 0-2-1, 2-1-0, 0-1-2, 1-0-1, 0-1-0, 1-0-0, and 0-1-1. System components were binned into profiles when a Pearson correlation coefficient > 0.88 was calculated. Only negligible changes in the assignment of profiles were found when using absolute or relative component units (e.g., mmol*gDCW-1vs. log2(FC vs ref)) or different correlation methods (i.e., Spearman).

System component statistical sample trend analysis:

Components identified from the differential tests (except for metabolomics) were used for sample trend analyses. Hierarchical clustering was used to diagnose sample groupings and distances between samples (distance metric of Euclidean and linkage method of complete). Principal component analysis (PCA) as encoded in the R package pcaMethods(Stacklies et al., 2007) (version 1.64.0, univariate scaling, centering, SVD PCA) was then used as a representative

unsupervised method to project samples into component space, and confirm the relative magnitude and direction of component weights. PCA models were first constructed for the reference, knockout, and endpoint for each of the lineages to confirm that the primary component best separated the reference and endpoint from the knockout, and that the second component best separated the reference and knockout from the endpoint. PCA models were then constructed for the reference, knockout, and all endpoints for each network perturbation. The PCA models were validated using cross validation (CV type of Krzanowski, default 5 segment with 5 CV runs per segment with minimum number of segments equal to the number of samples). Partial Least Squares Discriminatory Analysis (PLS-DA) was implemented using the R package pls(Mevik and Wehrens, 2007) (version 2.5, univariate scaling, centering, Canonical Powered Partial Least Squares (cppls) PLS-DA) was used to project replicate samples into component space. PLS-DA models were first constructed for the reference, knockout, and endpoint for each of the lineages to confirm that the primary component best separated the reference and endpoint from the knockout, and that the second component best separated the reference and knockout from the endpoint. PLS-DA models were then constructed for the reference, knockout, and all endpoints for each network perturbation. The PLS-DA models were validated using cross validation (default 10 segments with minimum number of segments equal to the number of samples).

The distance between the ref and uKO strain along axis 1 (i.e., mode 1) was used as a threshold to determine whether an eKO strain matched the

general mode 1 and mode 2 trends identified in section 2a. A relative distance for each eKO strain along axis 1 was calculated as follows: relative distance = distance(uKOj, eKOi,j)/distance(ref, uKOj) where i = endpoint replicate for a particular KO lineage and j = each KO lineage. An eKO strain with a relative distance greater than 70% along axis 1 was determined to match the trend.

Metabolite, flux, and gene set enrichment analyses:

Metabolite and gene set enrichment analyses were conducted using the subsystem categories of iJO1366. Flux and metabolite flux sum set enrichment analyses were conducted using the subsystem categories of iDM2015. A pvalue < 1e-3 (hypergeometric test) was used to test for enriched subsystems. Gene set enrichment analysis on differentially expressed genes was also performed using with R package topGO(Alexa and Rahnenfuhrer, 2010) with GO annotations for *E. coli* (Carlson, 2013). A p-value < 0.05 (Fischer statistic, parent-child algorithm(Grossmann et al., 2007)) was used to test for enriched biological processes and molecular functions.

Network distance and graph analyses:

The inverse mean values from sampled flux distributions that were constrained by the 95% confidence intervals derived from estimated MFA flux bounds were used as weights in calculating the shortest path from metabolite A to B. The iDM2015 network was deconstructed into a directed acyclic graph with metabolites and reactions composing the nodes and the connections between metabolites and reactions composing the links. Metabolites that did not contain carbon were excluded from the graph network. In addition, metabolites

corresponding to co2, co, mql8, mql8h2, 2dmmql8, 2dmmql8h2, q8, q8h2, thf, ACP were also excluded. Metabolites corresponding to udpglcur, adpglc, gam6p were substituted as glycogen_c, uacgam, uacgam, respectively, as they were not present in the lumped and reduced iDM2015 network. The A*star algorithm as implemented in the python package networkX (https://github.com/networkx/networkx) (version 1.11) was used to calculate the shortest path of the graph network. The distance from metabolite A to B was calculated as half minus 1 the computed shortest path.

A redistribution of flux was defined as a change in path or path length between the reference and knockout and endpoint or knockout and endpoint. A change in flux capacity was defined as a change in path or path length between the reference and knockout, but not between the knockout and endpoint.

Nodes (i.e., metabolites) were categorized as Intermediates, Carriers, Biomass Precursors, and/or Nucleotide Salvage Products as defined in Table S[]. The correlation (Spearman R, pvalue < 0.05) between path and path length and metabolite level was calculated between Intermediates and Carriers, Carriers and Biomass Precursors, Intermediates and Biomass Precursors, Carriers and Nucleotide Salvage Products, and Biomass Precursors and Nucleotide Salvage Products.

Multi-component -omics statistical analyses

Biomass to network component correlation analysis:

EcoCyc (Keseler et al., 2013) subsystems for the following biomass producing pathways were used in the analysis: Amines and Polyamines

Biosynthesis, Amino Acids Biosynthesis, Nucleosides and Nucleotides Biosynthesis, Fatty Acid and Lipid Biosynthesis, Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis, Cell Structures Biosynthesis, and Carbohydrates Biosynthesis. Gene identifiers from these pathways were mapped onto iDM2014 via the GPR relation to identify biomass producing reactions and metabolites. The analysis was conducted at the level of individual lineages using the system component profiles of restored-, novel+, overcompensation-, partially-restored-, and reinforced+ to identify positively correlated (correlation coefficient > 0.88, Pearson, r) with growth (i.e., growth promoting) and negatively correlated (correlation coefficient < -0.88, Pearson, r) with growth (i.e., growth inhibiting).

The number of significant biomass components were divided by the number of measured biomass components, and expressed as a percent. A direct pairwise correlation between metabolite concentrations, transcript levels, and fluxes, and growth rate was also performed (units of log2(FC vs. ref)) between the reference strain, knockout, and endpoints for all or each knockout condition for comparison (data not shown). Components that were positively correlated (correlation coefficient > 0.88, Pearson, r) with growth rate or negatively correlated (correlation coefficient > 0.88, Pearson, r) with growth rate were identified.

Inter- and intra- component correlation analysis:

A global pairwise correlation between metabolite concentrations, transcript levels, and fluxes was performed by comparing the agreement and disagreement between component profiles of restored+, novel+, overcompensation+, partially-restored+, unrestored+, and reinforced+. Components with matching profiles

with correlation coefficients > 0.88 (Pearson, R) were correlated; components with matching profiles with correlation coefficients < -0.88 (Pearson, R) were anti-correlated. A similar global pairwise correlation between metabolite concentrations, transcript levels, and fluxes was performed (units of log2(FC vs. ref)) for comparison (data not shown). Components with a correlation coefficient > 0.88 (Spearman, r) were correlated; Components with a correlation coefficient < -0.88 (Spearman, r) were anti-correlated.

Regulation to network component correlation analysis:

Significantly correlated components were compared to annotated gene-to-reaction, and metabolite-to-reaction interactions annotations in iJO1366, and to annotated transcription factor-to-gene, metabolite-to-transcription factor, metabolite-to-transcription factor-to-gene, metabolite-to-transcript, and metabolite-to-reaction regulatory interactions from the EcoCyc database (Keseler et al., 2013). EcoCyc database identifier were mapped to iJO1366 identifiers using a combination of ChEBI (Hastings et al., 2013), MetaNetX (Bernard et al., 2014; Ganter et al., 2013; Moretti et al., 2016), EC numbers, InCHi strings, and manual curation. The mode of component interactions were encoded as either positive for reactant-reaction, activating, or stabilizing interactions, or negative for product-reaction, inhibiting, or de-stabilizing interactions. The sign and magnitude of the correlation coefficient (Pearson, r) of matching categories was compared to the mode of interaction to determine agreement (correlation coefficient > 0.88 and positive mode, or correlation coefficient < -0.88 and negative mode). The inverse was used to determine disagreement. Similarly,

the sign and magnitude of the correlation coefficient (Spearman, R) was compared to the sign of the component interaction to determine agreement or disagreement for direct pairwise correlations (data not shown).

The classification of global regulators follows the definition given by Martinez-Antonio, et al., (Martínez-Antonio and Collado-Vides, 2003). Global transcription factors are defined to include CRP, IHF, FNR, FIS, ArcA, Lrp, and Hns. A secondary level of regulators are defined to include NarL, Fur, Mlc, CspA, Rob, PurR, PhoB, CpxR, and SoxR. The secondary level and lower level regulators (e.g., local transcription factors) were further broken into classes for local and general stresses.

Regulator activation categorization:

A profile for the activation status of each regulator for each knockout evolution was determined. The analysis was first limited to regulated entities that had only a single annotated regulator. The analysis was then expanded to include all regulators and regulated entities. A category weight for each regulated entity for each endpoint was calculated as follows: weight,i,j = abs(corr,i,j)*1/(nEPs,i)*1/(nRegulators,k) where i = endpoint, j = category, k = regulators, nEPs = number of endpoints per knockout evolution, corr = correlation coefficient, nRegulators = number of regulators per regulated entity. A confidence score for each regulator for each knockout was calculated as follows: confidence,i = sum(weight,i,j,k) where i = knockout, j = endpoint, and k = regulated gene. A higher confidence score indicates a consistently higher

correlation to the category across all regulated entities that are regulated by the regulator.

Results:

1a. Experimental design and nomenclature

The KO perturbation and recovery ALE experiment was performed in the model organism *Escherichia coli* K-12 MG1655. First, a starting strain in which the KO perturbations were implemented was chosen. A previously evolved wild-type strain was selected in order to isolate biological changes caused by adaption to the loss of a gene product from those caused by adaption to the growth conditions of the experiment. The pre-evolved strain (denoted "pre-evolved reference strain" or "Ref") was a previously described strain isolated from an ALE of wild-type *E. coli* evolved on glucose minimal media at 37°C (LaCroix et al., 2015). Ref was a non-mutator strain, had the fewest number of mutations of replicate ALE endpoints, and a relatively efficient conversion of glucose to biomass.

Second, perturbations consisting of five separate reaction KOs that were predicted to result in large metabolic rearrangements based on *in silico* analysis (see Methods, Table S1) were implemented in Ref. Genes (see Methods) encoding enzymes for the reactions of GND (*gnd*, 6-phosphogluconate dehydrogenase), GLCptspp (genes *ptsH*, *ptsI*, and *crr* corresponding to enzymes HPr, EI, and EIIA, respectively) SUCDi (genes *sucA*, *sucB*, *sucC*, and *sucD* corresponding to the enzyme Succinate Dehydrogenase), TPI (*tpiA*, Triosphosphate Isomerase), and PGI (*pgi*, Phosphoglucose Isomerase) were removed to generate strains uGnd, uPtsHIcrr, uSdhCB, uTpiA, and uPgi, respectively (denoted "unevolved knockout strains" or "uKO").

Third, replicates of the five knockout strains, as well as Ref, were simultaneously evolved on glucose minimal media at 37°C in an automated ALE platform (LaCroix et al., 2015; Sandberg et al., 2014) denoted “evolved knockout strains” or “eKOi” where i denotes the replicate number). The number of replicate endpoints were the following: 2 for “evolved reference strain” (denoted eRef), 3 for eGnd, 4 for ePtsHIcrr, 3 for eSdhCB, 4 for eTpiA, and 8 for ePgi.

Finally, the Ref, uKO, and eKO strains were subjected to multi-omics data generation under identical growth conditions. This data generation consisted of measuring intracellular metabolite levels, gene expression levels, flux levels, and genomic mutations (i.e., system components). Statistical and biochemical modelling methods were then applied to the measured data.

1b. Changes in fitness with enzyme KOs and ALE

A statistically significant loss and recovery of fitness (i.e., growth rate) was found in three of the KO strains. The initial fitnesses of uPgi, uPtsHIcrr, and uTpiAi were drastically reduced compared to the reference strain (81, 79, and 80% decrease in fitness, respectively) while the initial fitnesses of uGnd and uSdhCB were minimally changed (9 and 6% decrease in fitness, respectively) (Figure 1, Supplemental Data). A statistically significant increase in final fitness (Student’s t-test, pvalue<0.05) was found in all ALE endpoints of the ePgi, ePtsHIcrr, and eTpiA lineages (ave±stdev 284±20, 259±74, 164±7% increase in fitness, respectively) compared to the uKO for each lineage, while a non-significant and minimal increase in final fitness was found in all endpoints of the eRef, eSdhCB, and eGnd lineages (ave±stdev 4±1, 3±4, 5±3% increase in

fitness) compared to the reference strain or unevolved knockout strains (uSdhCB and uGnd lineages), respectively. The uKOs with the most drastically reduced initial fitnesses were not able to recover the fitness of the Ref, while all but one of the endpoints from the eSdhCB and eGnd lineages successfully recovered the fitness.

1c. Reference strain evolution confirmed the experimental design

An insignificant fitness change and the fewest number of network changes were found in eRef strains compared to all eKO strains (Table 1). The average numbers of significant component changes per eRef replicate at the metabolite, transcript, and flux levels were 2, 35, and 0, respectively. These changes in systems components were far fewer than in any of the other eKO strains, where the minimum number of corresponding changes were 19, 341, and 158. The average number of genomic mutations per eRef replicate was also the lowest of all lineages, and were primarily found in cell wall biosynthesis genes. Overall, these findings demonstrated that the use of a pre-evolved strain minimized the number of confounding component changes.

In the next subsection, the multi-omic data sets generated were analyzed at the *systems level* where the results from multi-dimensional analysis of the multi-omics data sets will be described and interpreted. Analysis at the *mechanistic level* follows where changes in eKO strains are detailed. Systems level analysis reveals an overall view of the adaptive evolution process, while analysis at the mechanistic level details how adaptation to the loss of a particular enzyme was achieved.

2a. Evolution to optimal fitness after gene KO was captured in the two dominant modes

The metabolite, transcript, and flux levels for each of the KO lineages were subjected to Partial Least Squares Discriminatory Analysis (PLS-DA) (see Figure 2 and Methods). For almost all cases analyzed, the first mode separated the reference strain and evolved knockout strains from the unevolved knockout strain, while the second mode separated the reference and evolved knockout strains (74% of eKOs from all data types and lineages, see Methods). In other words, the first mode accounted for a dominant transition between pre-optimized, perturbed, and re-optimized fitness states (i.e., captured systems fitness properties), while the second mode described alternate re-optimized states (i.e., capturing systems diversity, or a 'plateau' in the evolutionary landscape (Conrad et al., 2011)).

Two KO lineages that did not match the trend in the fluxomics data are worth highlighting. First, all eTpiA strains were not able to recover the distance to Ref along mode 1. This was primarily due to the bifurcation of flux usage in lower glycolysis after removing the *tpiA* gene that forced flux through the methylglyoxal detox pathway. Second, all eGnd strains were not able to recover the distance to Ref along mode 1. This lack of recovery was primarily due to a complete loss of the oxPPP after removing the *gnd* gene, and reversal of flux through the non-oxPPP in order to generate ribose for nucleotide synthesis.

2b. Profiles of changes in components reveals systematic variations between ALE lineages, KOs, and system components

In order to dissect the drive towards fitness (mode 1) and generation of diversity (mode 2) further, changes in each system component (i.e., metabolite, transcript, and flux level) between Ref, uKO, and eKO strains were grouped into six profiles (Figure S1A, see Methods): *novel, overcompensated, partially-restored, reinforced, restored,* and *unrestored*. The distribution between these six profiles for each component type are shown with horizontal bar charts in Figure S1B-D. Several trends were found based on these six profiles.

First, the occurrence of profiles varied between omics data types. Overall, the metabolite levels were the most distributed between the six profiles (i.e., had the least deviation). In contrast, the transcript levels were dominated by the Restored profile, and flux levels were dominated by the Restored and Unrestored profile. The more even metabolite distribution compared to the transcript levels or flux levels indicated that the changes in metabolite levels were less constrained to change than the gene expression and fluxes. Second, distribution amongst the profiles varied between KOs. The lineages with the greatest initial loss of fitness had a greater percentage of Novel, Overcompensated, Reinforced, and Unrestored profiles than the lineages with a smaller initial loss of fitness. This observation indicated that the larger the loss in fitness, the greater number of Innovative (as opposed to Restorative) network changes were required to regain fitness. Third, the distribution amongst profiles also varied between evolved strain lineages. This highlighted the biochemical differences in re-optimized network configurations during adaptation to overcome the perturbation. See Figure S1 for statistics and examples of these three trends.

The biological significance of the differing distributions in profiles between system components, KOs, and evolved lineages were further detailed in analyses and cases studies (Figures 3-6, Figures S2-7, and Tables S2-7) presented in the following subsections. The system component profiles are also used in all of the analyses presented below. The subsections (i-vii) develop a proposed model of adaptation to metabolic perturbation (Figure 7). This six-step model will be detailed with emphasis on the adaptation to the loss of PGI.

3i. Component imbalance retarded fitness

Alterations to biomass-producing pathways (e.g., amino acid biosynthetic pathways; See methods for all pathways, Figure S2, and Table S3) were examined in order to uncover the origins of the loss in fitness after the gene KO perturbations. Component profiles (Section 2, Figure S2) that increased during evolution were categorized as *growth-limiting* (i.e., correlated with growth), those that decreased during evolution were categorized as *growth-inhibiting* (i.e., anticorrelated with growth) (see Methods). Metabolic flux was found to consistently be the most growth-limiting across all lineages (Figure S2A-C). Gene expression was found to consistently be the most growth-inhibiting across all lineages. An imbalance in metabolite levels in biomass-producing pathways was also apparent.

Biomass-producing pathways that had more or less growth-limiting or -inhibiting components were identified (Figure S2D). These growth-limiting or -inhibiting pathways were found to align with the metabolic perturbation created. For example, the pgi lineages were metabolite-inhibited for nucleoside/nucleotide

biosynthetic pathways. The forced flux through the oxPPP after removing the *pgi* gene directed a disproportionate level of carbon into ribose-5-phosphate and towards de-novo purine/pyrimidine biosynthetic pathways. This resulted in abnormally high levels of L-histidine, IMP, and UMP in uPgi that were Unrestored, Restored, and Reinforced in the ePgi strains (Figure S2E). In a second example, the ptsHIcrr lineages were primarily inhibited by metabolite excess for the amino acid biosynthetic pathways. The elevated levels of phosphoenolpyruvate after removing the ptsHIcrr genes resulted in an abnormally high level of aromatic amino acids compared to other amino acids in uPtsHIcrr that were Restored in most ePtsHIcrr strains(Chávez-Béjar et al., 2012/7; Flores et al., 1996) (Figure 4C). See Figure S2 for additional examples.

This analysis revealed that the origins of the loss in fitness following perturbation were caused by an imbalance of system components that propagated to an imbalance in biomass-producing pathway usage. The perturbation-specific response for each of the KO lineages indicated a unique set of growth limitations and inhibitions.

3ii. Suboptimal pathway usage limited allocation of carbon to biomass precursors.

Above, limitations in metabolic flux was found to be the most growth-limiting system component. In order to understand the effects of metabolic flux limitation on fitness, changes in pathway usage between the Ref, uKO, and eKO strains were calculated, and grouped into *changed flux distribution* (i.e., the pathway usage was changed) or *changed flux capacity* (i.e., the same pathway

was used but at a higher flux level, see Methods for extended definitions, Figure S3A-D).

Changed flux distribution was found to be more prevalent than changed flux capacity. Changed flux distribution was found to occur 55.6% of the time, while a change in flux capacity was found to occur 22.0% of the time across all perturbations and lineages (Figure S3E). The remaining 22.4% of cases were unaffected. The ratio between the occurrence of changed flux distribution and changed flux capacity was consistent for the pgi, gnd, and ptsHIcrr lineages, but differed for sdhCB (84.5% changed flux distribution and 5.0% changed flux capacity) and tpiA (35.8% flux distribution and 26.6% changed flux capacity). This result indicated that the expressed enzymatic machinery post-knockout both proximal and distal to the network lesion was sub-optimally suited to distribute flux optimally towards biomass precursors.

Several examples of changed flux distribution and capacity are worth highlighting. In uPgi, the abnormally high levels of flux directed through the oxPPP was initially re-routed through the ED pathway (Figure S3I). Several ePgi strains retained the flux through the ED pathway to varying degrees, but most ePgi strains re-distributed flux through GND, and all increased the flux capacity through the non-oxPPP. In another example, flux was initially re-routed through the ED pathway in uGnd (Figure S3F) in order to generate ribose through the non-oxPPP. The ED pathway has a net yield of one ATP, NADH, and NADPH per molecule of glucose, whereas glycolysis has a net yield of two ATP and NADH (Nelson and Cox, 2013). Instead, the eGnd strains limited the

use of the oxPPP and increased the flux capacity through the higher energy and redox equivalent producing pathway of glycolysis. Further examples are given in Figure S3.

These results indicated that the initial flux distribution of the uKO strains following perturbation were often suboptimal, and required a change primarily in flux distribution and secondarily in flux capacity in order to restore fitness in the eKO strains. The re-distributed and increased capacity of pathways contributed to a reallocation of carbon towards biomass precursors in the ratios that were required to recover optimal fitness.

3iii. Perturbed metabolite levels triggered transcription regulatory network responses in uKOs.

Perturbed metabolite levels were traced to known transcriptional regulatory network (TRN) responses (Cho et al., 2011a, 2015; Federowicz et al., 2014; Gama-Castro et al., 2016; Kim et al., 2012). Measured metabolite profiles were mapped to metabolite-activated TFs by comparing the relationship (i.e., positive or negative) between a metabolite profile, a TF that interacts with the metabolite, and the expression profiles of the TUs regulated by the TF (see Methods, Table S5).

Strong evidence for changed TF activation profiles (analogous to the system component profiles, Figure S1) were identified for 75 TFs (Table S6, Figure S4). These included 7 global TFs (i.e, CRP, Fis, IHF, ArcA, Lrp, FNR, and HNS (Martínez-Antonio and Collado-Vides, 2003)) and 68 local TFs (see Methods). The activation profiles of 15 TFs (which included the 7 global TFs,

and the 8 local TFs ArgR, CpxR, Cra, Fur, NsrR, OxyR, PhoB, and TyrR) were changed across all lineages. The remaining 60 TFs appeared to be changed in a perturbation and lineage-specific manner.

Interestingly, TF activation and their gene expression was not coincidental (ave±std 5.4±3.8, 4.1±2.6, and 70.5±6.1% agreement, disagreement, no significant change in expression profile per lineage, respectively, Table S4). This result indicated that the majority of changed TF activation profiles were attributed to changed concentrations in their activators (e.g., through small-molecule binding) as opposed to changed TF gene expression levels. Similar observations have been made for sigma factors and the expression levels of their regions in response to a key rpoB mutation(Utrilla et al., 2016). Several examples of changed global and local TF activation profiles by changed metabolite concentrations are given below.

The strongest example of a change in global TF activation was that for CRP. CRP regulates hundreds of genes involved in a multitude of processes including alternate carbon metabolism(Deutscher, 2008), osmoregulation(Balsalobre et al., 2006), biofilm formation(Jackson et al., 2002), etc., and has been shown to coordinate optimal proteome allocation for different nutrient conditions based on the levels of cAMP(You et al., 2013). A changed CRP activation was found in all lineages due to elevated levels of cAMP in the uKOs (Gunasekara et al., 2015). CRP was not differentially expressed in any of the lineages, but Restored cAMP levels were mirrored by Restored gene

expression of transcription units (TUs) solely regulated by CRP-cAMP (Table S6).

ArcA provided another example for global TF activation without a significant gene expression change. The restored activation profiles of ArcA (Alvarez and Georgellis, 2010) and several other iron-sulfur cluster homeostasis TFs found in all lineages could be linked to changes in TCA cycle intermediates as well as quinone pools (e.g., gnd and sdhCB, Figure 6). The arcAB two-component system in particular modulates genes in response to changes in respiratory conditions that are communicated via the intermembrane quinone pools.

Local TF activation was also identified in the uKOs. A change in activation of the PurR regulator was found in pgi and several other lineages due to changed levels of purine degradation products. Specifically, the *purR* dimer binds hypoxanthine and guanine, and regulates genes involved in purine metabolism (Cho et al., 2011b; He et al., 1990; Meng et al., 1990). The concentration profiles of hypoxanthine and/or guanine matched the expression profile for purR-target genes, while the expression profile for *purR* itself did not (Table S5).

Another example of local TF activation involved the use of small regulatory RNA. Abnormal elevations in glucose 6-phosphate (g6p) and an imbalance of the glycolytic intermediates in uPgi were found to induce a sugar phosphate toxicity response sensed through the TF SgrR and mediated through the action of the small RNA *sgrS* (Richards et al., 2013; Vanderpool and Gottesman, 2004, 2007). SgrR is thought to bind hexose phosphates and induce the expression of *sgrS*

(Figure S7). It was found that the metabolite concentration profiles matched *sgrS* expression profiles. *SgrS* transcriptionally regulates a number of genes that are involved in re-balancing glycolytic intermediates. One target of *sgrS* attenuation is *purR*, which explains the opposing *purR* expression profile compared to its TF activation profile described above. Interestingly, abnormal elevations of g6p and induction of SgrR and SgrR regulons were also found in ptsHIcrr. Additional examples are provided in figure S4.

Thus, the individual lineages provided clear examples of global and local TF activation through changes in activating metabolite concentrations through mechanisms that are well-established in the literature. The TFs examined here were associated with carbon metabolism, nitrogen metabolism, iron regulation, oxidative stress, DNA repair, and other stress responses that control the majority of known genes in *E. coli*.

3iv. Transcription factor responses resulted in a misallocation of resources or amplification of processes reducing fitness in uKOs.

The global and local TF responses that were triggered by changed metabolite concentrations in the uKOs strains were investigated further to assess their impact on the observed changes in fitness. A common theme that emerged was that many TF responses resulted in the upregulation of metabolic pathways and biological functions that were counterproductive to fitness. Such adverse responses showed that regulatory circuits in the uKO strains were no longer ‘tuned’ to support fitness of the metabolic network after gene KO.

One example of a counterproductive TF response was the activation of CRP by elevated cAMP levels. Such CRP activation generated a cascade of translation in alternate carbon metabolism operons (Hermsen et al., 2015; You et al., 2013). These operons pertained to import and catabolism of sugars such as glycerol, maltose, mannose, etc., that were not present in the medium (Supplemental data). Specifically, the *glp* regulon required for glycerol import and catabolism is up regulated by CRP-cAMP (Larson et al., 1992). This hard wired regulation led to massive up-regulation of the *glp* regulon in uPgi leading to unproductive allocation of the proteome to glycerol metabolism.

In another example, the glucose-6-phosphate (g6p) concentration build-up in uPgi was so great that g6p spilled over into the periplasmic space (Bolten et al., 2007; Link et al., 2008) (Figure 3, upper panels). High periplasmic g6p was sensed by the uhpAB two-component system, which in turn up-regulated expression of the hexose phosphate importer *uhpT* (Dahl et al., 1997; Maloney et al., 1990; Weston and Kadner, 1988). Increased expression of *uhpT* likely generated a loop whereby excessive g6p that spilled into the periplasmic space would be re-imported into the cytosol.

These examples illustrate how the loss of an enzyme led to an unorganized, confused, and suboptimal allocation of resources by inducing previously evolved regulatory responses. As shown below, these hardwired responses were modulated and re-wired during adaptation to achieve better fitness.

3v. Alternate regulatory mechanisms abrogated counterproductive transcription factor responses.

Cells contain multiple levels of counteracting regulatory mechanisms. Regulatory mechanisms at the flux level were identified that partially counteracted the adverse response to enzyme KO at the expression level (i.e, the activating TF response discussed above). The agreement and disagreement between the system component profiles (described in Section 2b) and known biochemical pathways (Orth et al., 2011) and their regulation (Keseler et al., 2013) was determined (see Methods, Figure S4). A relatively low agreement between changes in gene expression profiles and metabolic flux profiles (i.e., gene-protein-reaction association, GPR) within each lineage was found (Table S4). Specifically, an average agreement of 27.5% (stdev=17.4%, n=22) and average disagreement of 11.5% (stdev=6.8%,n=22) was found. A similarly low agreement between types of literature-derived regulation were found (see Methods, Table S4).

It was hypothesized that the low agreement found between genes and fluxes and between system component profiles and known regulation reflected either 1) counteracting regulatory mechanisms, 2) evidence for inaccurate or incomplete knowledge of the regulatory network (Covert et al., 2004; Koussounadis et al., 2015; Maier et al., 2009; Vital-Lopez et al., 2013), or 3) changes to regulation introduced through fixed mutations. Evidence of competing layers of regulation for 89 regulators (i.e., any biological component that can effect a change in another component, e.g., TF or small-molecule)

across 5887 regulated entities (i.e., any biological component that is subject to regulation, e.g., TU or enzyme) were found. Evidence of inaccurate or incomplete knowledge of the regulatory network in 38 regulators across 631 regulated entities were also found (Table S6). While it is infeasible to investigate each discrepancy here, specific examples are given that illustrate the above three hypothesized mechanisms.

In an example of counteracting regulatory mechanisms, a hierarchy of transcription factor control over gene expression was recapitulated. The activation profile of Fis (Bradley et al., 2007; Cho et al., 2008; Weinstein-Fischer and Altuvia, 2007) was found to conflict with its consensus activation profile of the *pyrD* promoter in all of the pgi lineages, whereas the PurR activation profile was found to agree with pyrD expression profile (Bradley et al., 2007; Cho et al., 2008, 2011b). This indicated that *pyrD* expression was dominated by PurR regulation. In another example, a restored activation of *sgrS* found in the pgi lineages and a novel activation of *sgrS* found in the ptsHIcrr endpoints 1 and 3 negated the transcription factor regulation of sgrS target genes (Bobrovskyy and Vanderpool, 2016; Sun and Vanderpool, 2013). Further examples are given in (Figure S4).

Unresolved discrepancies in regulatory annotations were found. The expression profiles of regulons that were controlled only by Fur (Beauchene et al., 2015; Chen et al., 2007; Méhi et al., 2014) were found to be inconsistent. Specifically, the expression profiles for *entS*, *exbB*, *exbD*, *fecI*, *fepC*, *fepD*, *fhuA*, *fhuE*, *ryhB*, and *yjjZ*, conflicted with that of *crl*. The discrepancies indicated that

another TF or transcriptional regulator is present that also controls the transcript levels of that gene or Fur can act as a dual regulator similar to entS (Lavrrar et al., 2002). In fact, *crl* has been shown to also be regulated by ArgR (Cho et al., 2011a) and positively regulated by CsrA(González Barrios et al., 2006). In addition, *yjjZ* has also been shown to be positively regulated by OxyR (Seo et al., 2015a) and positively and negatively regulated by Fnr(Federowicz et al., 2014).

Discrepancies arising from changes to regulation introduced through mutation were also identified. For example, the *lon*-specific promoter is activated by GadX (Gama-Castro et al., 2016; Seo et al., 2015b; Tramonti et al., 2008). A mutation at the *lon*-specific promoter in the ePgi replicates 1-5 silenced the expression of *lon* thereby negating the regulation by GadX (Figure S7). This silencing directly affected the expression of colanic acid and biofilm producing operons that are controlled by RcsA and RcsAB (Majdalani and Gottesman, 2005). The Lon protease degrades RcsA (Torres-Cabassa and Gottesman, 1987).

The examples given above indicate that the response of the uKO and eKOs recapitulated the effects of known regulation, but also revealed the effects of unknown or not fully characterized regulatory mechanisms. The latter provide suggestions for new experimental lines of inquiry. The examples of overlapping and competing layers of regulation also help explain and identify built-in regulatory mechanisms that mitigated the counterproductive regulatory responses. Adaptation and reconfiguration of regulatory mechanisms caused by mutations fixed during ALE are given in the next subsection.

3vi. Mutations selected during adaptive evolution re-wired many counterproductive system responses.

A large number of mutations were identified in the eKOs that either offset the counterproductive effects of global and local regulators (discussed above) or targeted specific pathways or imbalances. In total, 673 mutations were found in the eKOs. The mutations were found to primarily be single nucleotide polymorphisms (SNPs, 66%), were primarily located in coding regions (48%), and were primarily associated with membrane proteins and transcription factors (27 and 29%, respectively). See Table S7, and Figure S5 for a detailed overview of all genomic mutations found in the eKO strains.

Mutations were identified that offset the mis-regulation of global TFs. For example, a substantial number of mutations affected regulators of carbon transport and metabolism processes that appear to offset the activation of operons induced by CRP-cAMP. These included mutations to *galR*, *malT*, and *crr* in the ePgi strains (Figure S6). A 22 nucleotide deletion in the small molecule binding domain of *galR* in ePgi07 appears to negate repression of *galR* controlled operons. These include *galETKM*, *galP*, and *mglBAC* that encode enzymes for galactose catabolism, symport, and ABC transport, respectively(Weickert and Adhya, 1993). These operons are also regulated by CRP-cAMP, and were not expressed in Ref. The galactose importers have lesser affinity for the transport of glucose, which may give ePgi07 an additional route to import and catabalize glucose from the environment. In addition, the mutation may have aided in conserving PEP for aromatic amino acid production,

which was limiting fitness in all of the pgi strains (discussed in section 3i). Interestingly, mutations in *galR* or at the *galR* operon in ePtsHIcrr02/04 and in eTpiA01/03 also resulted in the upregulation of GalR controlled genes. The prevalence of *galR* mutations may indicate that expression of the *gal* regulon may aid in increasing fitness when the ability to import glucose is impaired or the levels of PEP are inadequate for aromatic amino acid production. Additional mutations that affected carbon transport processes included *ptsG*, *galR,* and *nagC* in the ePtsHIcrr strains (data not shown), and *ptsG*, *galR*, and *nagA*, *nagC*, and *nagE* in the eTpiA strains (data not shown).

Mutations were also identified that offset the mis-regulation of local TFs. For example, the methylglyoxal pathway in tpiA was tuned to more efficiently convert methylglyoxal to lactate through mutations that altered methylglyoxal detox pathway gene expression. There are four routes in *E. coli* by which methylglyoxal can be metabolized; one route involves the *gloAB* enzymes that utilizes glutathione to convert methylglyoxal to lactate (Figure S8). Expression of *gloA* is repressed by NemR; repression is enhanced by elevated levels of methylglyoxal (Ozyamak et al., 2013; Umezawa et al., 2008). A mutation in the *nemR* promoter region in eTpiA03, and mutations in the *nemR* small-molecule-binding domain or tetramerization regions in eTpiA01, 02, and 04 were found to offset NemR repression, and allowed for increased expression of gloA (Figure 5). The increased expression of *gloA* appeared to have provided a fitness advantage by increasing the conversion of DHAP to Lactate and subsequent conversion of Lactate to Pyruvate (Figure 5). The levels of Pyruvate were severely depleted in

uTpiA due to the forced bifurcation of flux that resulted from the loss of the triosephosphate isomerase enzyme.

3vii. Mutations selected during adaptive evolution also introduced innovations that targeted specific pathway or metabolite imbalances.

Mutations were identified that targeted specific pathway or metabolite imbalances. One of the clearest examples of this was in the pgi strains, where abnormally high flux levels through the oxidative pentose phosphate pathway resulted in the overproduction of NADPH. Overproduction of NADPH was first buffered by glutathione (Supplemental data) and further mitigated through mutations in isocitrate dehydrogenase (Figure 3) and the transhydrogenases (Supplemental data). A point mutation at the 395 residue that changes the amino acid from positively charged (L-arginine) to negatively charged (L-cysteine) in isocitrate dehydrogenase was found in all ePgi replicates except replicate 7. The mutation occurs 4 Angstroms from the phosphate moiety of NADP. The 395 residue has been shown to be directly involved in NADP-binding (Zhu et al., 2005), and appears to allow the mutated enzyme to utilize NAD as a cofactor. The mutation was found to redirect flux through the glyoxylate shunt instead of the TCA cycle, and may provide a fitness advantage to the ePgi strains by limiting the production of NADPH in the TCA cycle.

In another example of mutations that targeted specific imbalances, the ATP drain caused by the use of alternative glucose importers in ptsHIcrr resulted in a significant reduction in the energy charge, and the decreased availability of energy equivalents contributed to a loss in fitness (Balderas-Hernández et al.,

2009; Fuentes et al., 2013; Valgepea et al., 2011) (Figure 4). The loss of the ptsHIcrr genes also resulted in a lexA-mediated SOS response (Kreuzer, 2013) (data not shown). The SOS response upregulated a plethora of DNA repair genes as well as genes known to confer a mutator phenotype. The lexA-mediated SOS response manifested in a large duplication that included the ATP synthase complex in ePtsHIcrr02/04 (Figure 4). The increased gene dosage of ATP synthase complex genes most likely contributed to the significant increase in the energy charge that was found in ePtsHIcrr02/04.

Discussion:

Although mutations can be found and causality established at the genome-scale (Herring et al., 2006), the mechanisms and principles that underlie adaptation and laboratory evolution have not been revealed. Here, the combination of study design, automated ALE, multi-omic data sets, and statistics and bioinformatics revealed general, KO-specific, and lineage-specific mechanisms of adaptation. These are detailed in the case studies presented in the text, as well as in the main and supplemental figures. The three common principles of adaptation were revealed (Figure 7B) that build upon previous work that have investigated ALE (Carroll and Marx, 2013; Charusanti et al., 2010; Cooper et al., 2003, 2008; Gresham et al., 2008; Kvitek and Sherlock, 2011; Lenski et al., 2015; McDonald et al., 2009; Toprak et al., 2011). They represent a first step towards developing a fundamental understanding of how cells *mechanistically* adapt to environmental and genetic change from a *systems* perspective that goes above and beyond general concepts of variability, heritability, and reproduction.

The results of this study led to a model of systems adaptation (Figure 7A) whereby imbalances in metabolite levels from altered fluxes triggered a multitude of network responses that were readjusted by mutations selected for during evolution. The mutations that fixed during ALE acted to rewire many existing hardwired responses and/or introduce novel network functions that addressed the imbalances that the initial KO lesion created. Given the results of KO-specific case studies, caution should be taken when interpreting the results of single

gene KO experiments because it is difficult to assess causal network changes without considering the adaptive network response to the KO. Novel mechanisms and inconsistencies, revealed through ALE, between measurement and known regulatory mechanisms identified in the case studies present opportunities for future discovery.

Taken together, this study highlighted the need for an approach whereby genes and cellular components were not analyzed in isolation, but instead where genes, cellular components, and their interactions were analyzed in the context of the cells' physiological functions. The genetic perturbations made in this study and the subsequent ALE experiments represent clear examples of the importance of a systems perspective in understanding optimization and re-optimization of cellular functions, and how cellular components, often surprising ones, must be adjusted to achieve optimal fitness. The study re-enforces the wisdom of the well known quote: "Nothing in biology makes sense except through the eyes of evolution" -- Theodosius Dobzhansky.

Acknowledgments

Chapter 8, in full, is a reformatted submission of “Laboratory Evolution of Gene Knockout Strains Reveals Fundamental Principles of Adaptation.” McCloskey D., Xu S., Sandberg T.E., Brunk E., Hefner Y., Szubin R., Feist A.M., and Palsson BO. Cell. The dissertation/thesis author was the primary investigator and author of the paper.

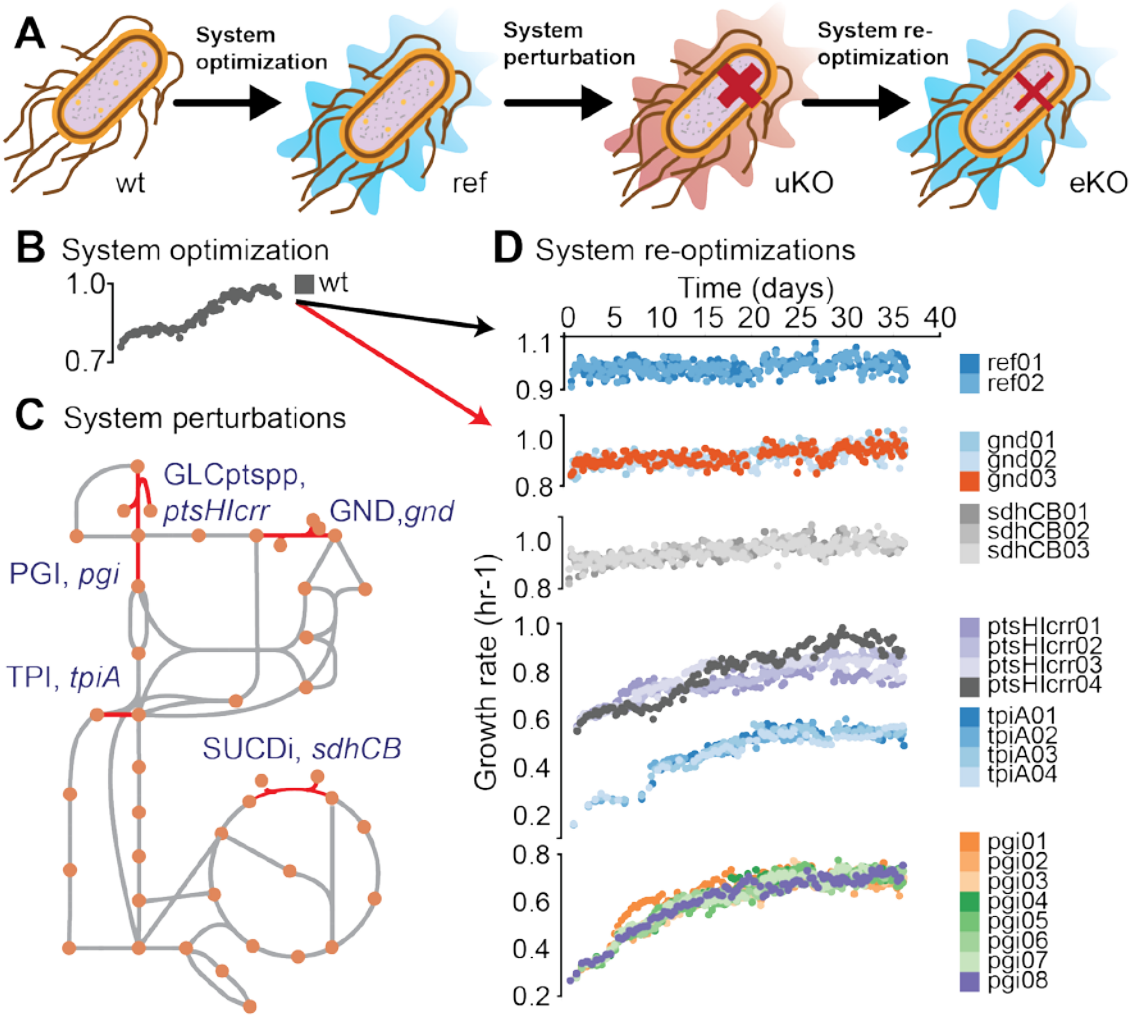

Figure 8.1: Evolution of knock-out strains from a pre-evolved (i.e. optimized) wild-type strain. A) Experimental design using adaptive laboratory evolution (ALE) and enzyme knockouts to investigate system re-optimization following major metabolic perturbations. B) Wild-type (wt) *E. coli* (MG1655 K-12) was previously evolved on glucose minimal media at 37°C (LaCroix et al., 2015). An isolate from the endpoint of the evolutionary experiment was selected as the starting strain for knockouts of key metabolic genes and subsequent re-evolution, or systems re-optimization. C) Reactions disabled by the enzyme knockouts included the phosphotransferase sugar import system (ptsHIcrr), phosphoglucose isomerase (pgi), 6-phosphogluconate dehydrogenase (gnd), triophosphate isomerase (tpi), and succinate dehydrogenase complex (sdhCB). D) Adaptive laboratory evolution trajectories of the initial reference knockout (KOs) and evolved knockout lineages. E) Counts of significantly different system components found for each evolved knockout relative to the unevolved knockout. Counts of metabolomic, transcriptomic, and fluxomic data are given as the average and standard deviation of the percent of significant features compared to all features measured for the lineage; counts for mutations are given as the average and standard deviation of the number of significant features (See Methods for criteria for significance).

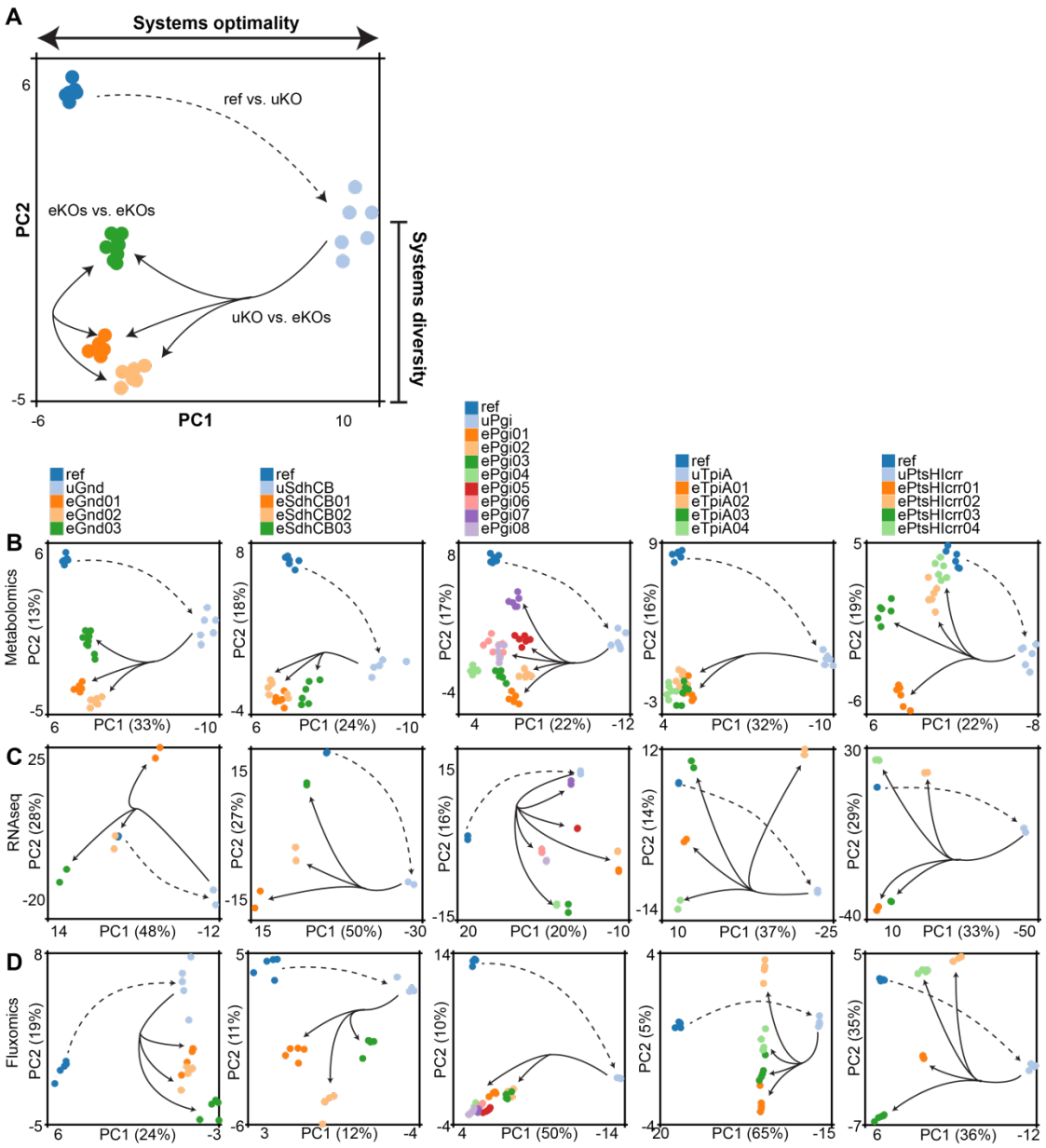

Figure 8.2: A multivariate analysis of biological network components as represented by different omics data types. A) Partial Least Squares Discriminatory Analysis (PLS-DA) revealed a common trend in the two most dominant components: the primary component (PC1) most often corresponded to a movement away from (dashed line) and back to (solid line) evolved optimal fitness (i.e., optimal system configuration), while the secondary component (PC2) most often corresponded to a diversity among evolved optimal fitness states of different lineages (i.e., optimal system configurations). PLS-DA scores plots of the reference strain, initial knockout, and evolved endpoints for each lineage for metabolomics (B), transcriptomics (C), and fluxomics (D) data. The strain lineages denoted on the top of panel B also refer to the corresponding graphs below in Panels C and D. All of the KO lineages matched the trend described above in the metabolomics data, one eKO did not match the trend in four of the five KO lineages in the expression data (i.e., all but eSdhCB), and one or more eKO did not match the trend in each of the KO lineages in the fluxomics data (see Methods for thresholds).

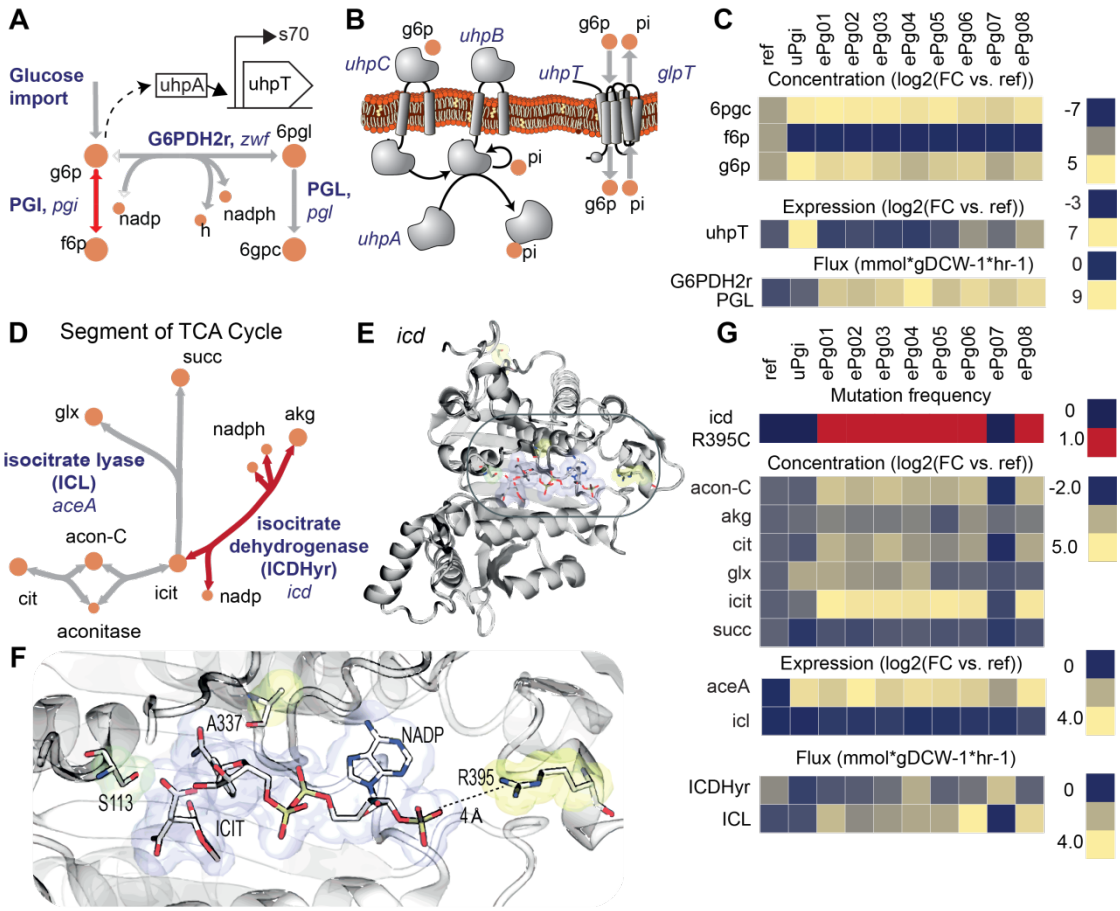

Figure 8.3: Proximal and distal network responses to the loss of phosphoglucose isomerase (PGI). Proximal: Knockout of PGI led to a hexose phosphate (panels A-C) and redox imbalance (not shown). As an example, the magnitude of g6p in the initial knockout led to a deleterious cycle whereby leakage of hexose phosphate across the inner membrane (Bolten et al., 2007; Link et al., 2008) induced hexose phosphate re-uptake via the uhpBC two-component system and uhpT hexose phosphate transporter (Dahl et al., 1997; Maloney et al., 1990; Weston and Kadner, 1988). Other examples of the consequences of the PGI KO are provided in the Supplement. A) A network map and regulatory schematic of the reactions into and out of the g6p node. The reaction in red is removed through the PGI KO. B) A mechanistic schematic of the uhpBC two-component system that sensed periplasmic hexose phosphate. The transcription factor UphA positively upregulated the expression of the hexose phosphate importer uhpT. C) Metabolite, expression, and flux levels near the node of perturbation. Distal: A beneficial mutation that rewired the TCA cycle via a cofactor usage swap in isocitrate dehydrogenase (ICD) aided in alleviating the excessive conversion of NADP to NADPH (Panels D-G). D) Network schematic of a segment of the TCA cycle. The reaction in red is catalyzed by ICD where mutations fix during ALE. E) Crystal structure of ICD. The mutated amino acids are highlighted in yellow. F) Zoom in on the active site of isocitrate dehydrogenase showing the proximity of the mutated amino acid to the phosphate group of NADP. The mutation occurs 4 Angstroms from the phosphate moiety of NADPH. The 395 residue has been shown to be directly involved in NADPH-binding (Zhu et al., 2005), and appears to allow the mutated enzyme to utilize NADH as a cofactor. G) Mutation frequency and metabolite, expression, and flux levels near the mutated gene. System components near the ICDHyr reaction in the ICD mutant strains are significantly changed.

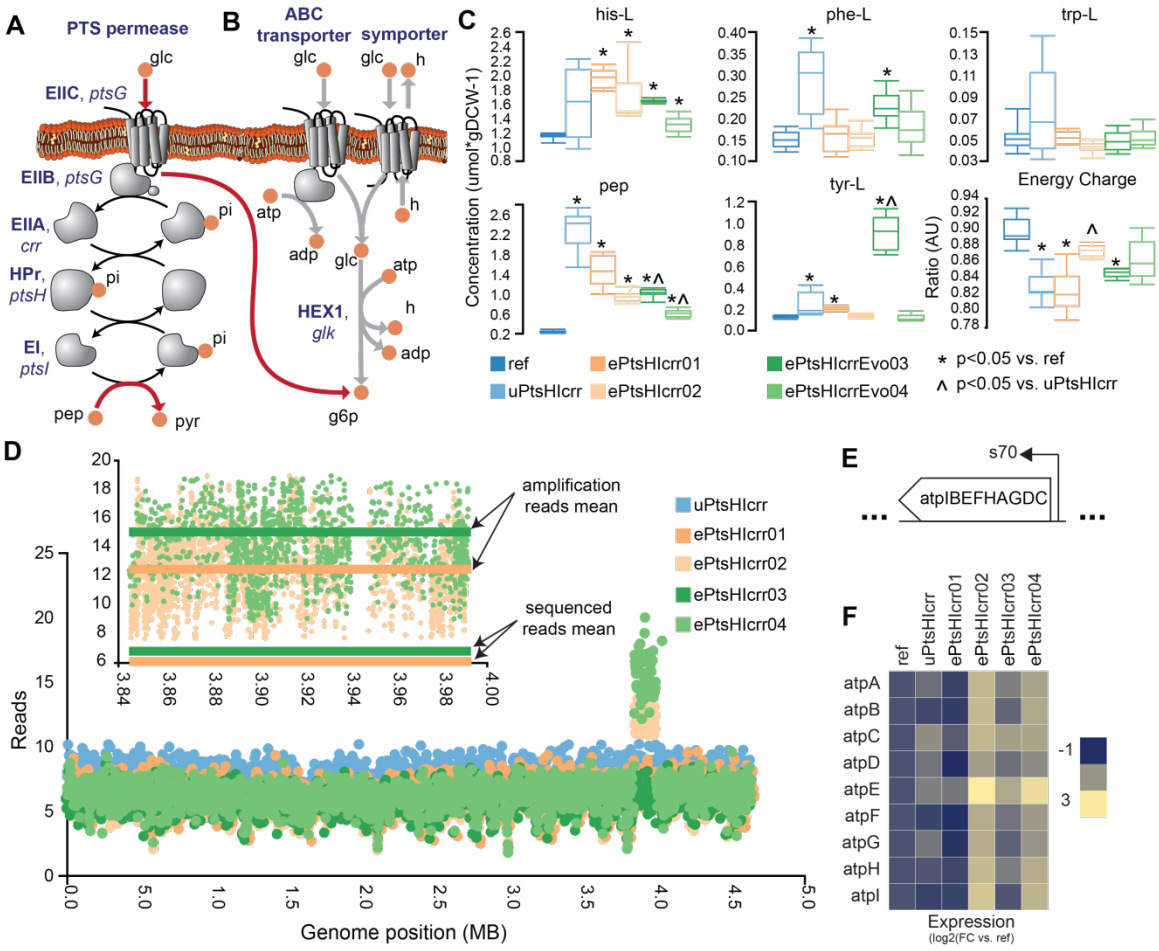

Figure 8.4: Proximal and distal network response loss of the Phosphotransferase System (PTS). Proximal: Knockout of the primary glucose importation system (ptsHIcrr) increased the availability of PEP for aromatic metabolite production, but caused a drain in available ATP through the upregulation of secondary glucose importation systems that utilize ATP instead of PEP (Panels A-C). A) A network and mechanistic schematic of the PTS system. The metabolite conversions in red are removed through the ptsHIcrr KO. B) A schematic of passive/active glucose importers. C) Metabolite concentrations of the aromatic amino acids L-phenylalanine (phe-L), L-tryptophan (trp-L), and L-tyrosine (tyr-L), and their precursor phosphoenolpyruvate (pep). Also shown are the metabolite concentrations for L-histidine, which is derived from ribose, and energy charge ratio calculated as (atp + adp/2)/(atp +adp + amp). Distal: Knockout of ptsHIcrr also induced a lexA-mediated SOS response (Kreuzer, 2013) (not shown). This manifested into a large chromosomal duplication event that resulted in an increased gene dosage of ATP synthase complex genes that most likely aided in restoring the energy charge (Panels D-F). D) Reads vs. genome position. Inset highlights the duplicated region near 4MB. E) Schematic of the ATP synthase operon genes. F) Expression levels of the ATP synthase complex genes. ePtsHIcrr02/04 ATP synthase genes are significantly elevated; note that the energy charge for ePtsHIcrr02/04 is not significantly different than Ref.

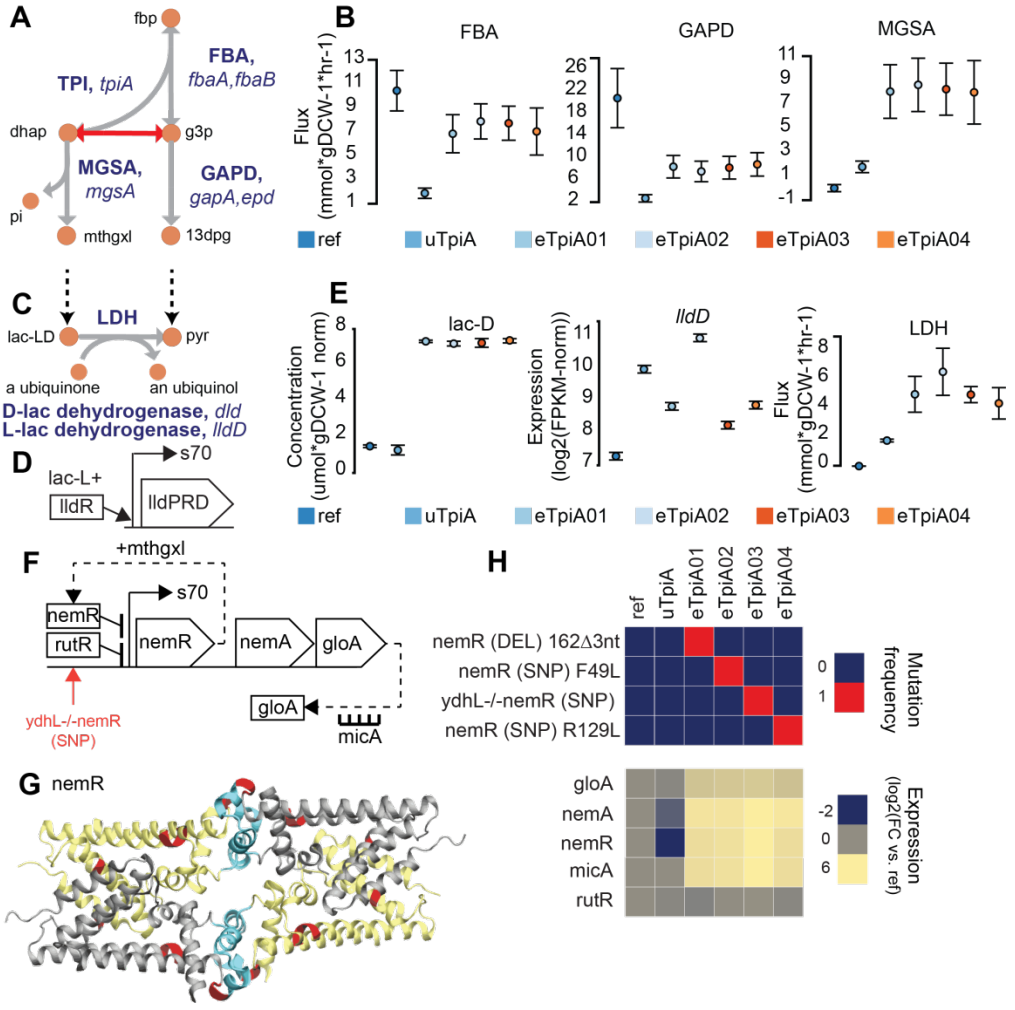

Figure 8.5: Proximal and distal network response to loss of triose phosphate isomerase (TPI). Proximal: The removal of TPI forced a bifurcation of flux in glycolysis through a pathway that involved the toxic intermediate methylglyoxal (mthgxl) (panels A-E). The bifurcated flux is rejoined by the lactate dehydrogenase (LDH) reaction. A) A network schematic of middle glycolysis. Shown along with removed TPI (shown in red) are the reactions catalyzed by fructose bisphosphate aldolase (FBA), glyceraldehyde 3-phosphate dehydrogenase (GAPD), and methylglyoxal synthase (MGSA). B) Flux levels of reactions near the gene knockout. C) Network schematic of lactate and pyruvate conversion. D) Regulatory schematic showing the feedback loop that up-regulates the conversion of D/L-Lactate to pyruvate when elevated levels of intracellular lactate were sensed by the transcription factor lldR. E) Metabolite, expression, and flux levels involved in the regulatory feedback loop. Distal: Mutations that affect the transcription factor NemR that allowed for expression of the gloA detox pathway(Ozyamak et al., 2013) (Panels F-H). F) Schematic of the nemRA-gloA operon (see Figure S8 for expression profiles of the methylglyoxal detox pathways). NemR exerts negative feedback on the operon that is enhanced by increased levels of methylglyoxal (mthgxl) (Ozyamak et al., 2013; Umezawa et al., 2008). gloA is co-expressed with nemR and nemA genes. A mutation that alters NemR binding to the regulatory region is annotated in red. G) Crystal structure of NemR (Gray et al., 2015). Chains A, B, C, and D are highlighted in gold and grey; the regulatory region is highlighted in cyan; mutations are annotated in red. H) Mutation frequency and expression profiles of nemR and nemR-associated genes. The lineages with mutations have increased expression of the gloA detox pathway.

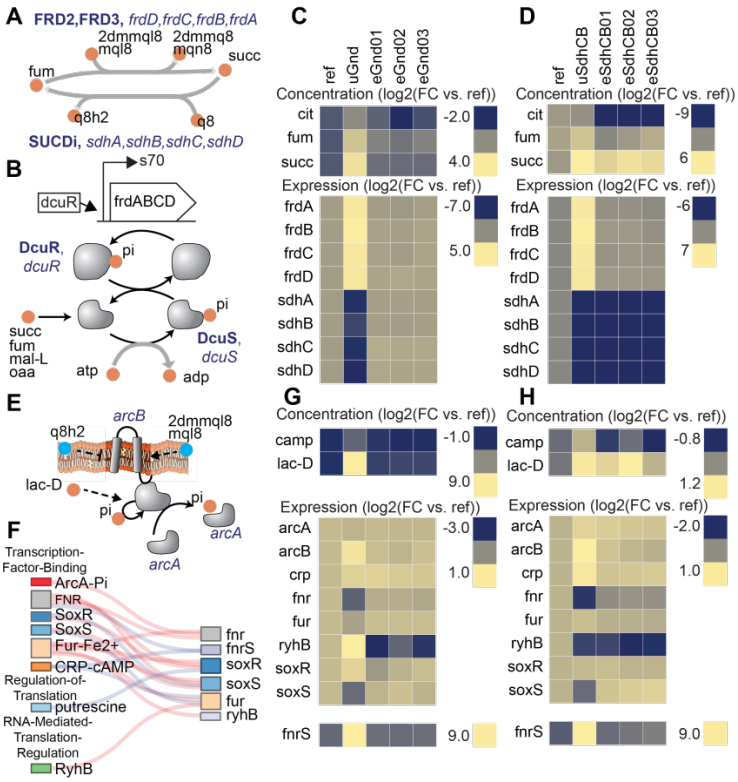

Figure 8.6: Perturbations in separate network locations yielded similar expression states in the uKOs due to similar metabolite levels. Removal of the succinate dehydrogenase complex (sdhCB), which decoupled the TCA cycle from oxidative phosphorylation, and removal of 6-phosphogluconate dehydrogenase (gnd), which re-routed flux through upper glycolysis, the ED pathway, and the pentose phosphate pathway (PPP), resulted in similar expression profiles of TCA cycle genes as a result of increased levels of intracellular four-carbon acids (Panels A-D). A) Reactions catalyzed by succinate dehydrogenase (SUCDi) and fumarate reductase (FRD2, FRD3) in the TCA cycle. B) Schematic of the frd operon and regulation by DcuR. Also shown is a schematic of the dcuRS two component system(Janausch et al., 2004). Elevation in four-carbon acids (e.g., succinate, fumarate, malate, and oxaloacetate) were detected by the dcuRS two-component system in the uKO strains. Phosphorylated DcuR activated expression of the fumarate reductase operon genes. Metabolite levels of fumarate (fum), succinate (succ), and citrate (cit), and expression levels of fumarate reductase (frdA, frdB, frdC, frdD) and succinate dehydrogenase (sdhA, sdhB, sdhC, sdhD) genes for gnd (Panel C) and sdhCB (Panel D). The similar metabolite levels in the uGnd and uSdhCB activated a network response that resulted in the upregulation of frdABCD genes in uGnd and uSdhCB, and downregulation of sdhABCD genes in uGnd. De-coupling of the TCA cycle from oxidative phosphorylation triggered an attenuated anaerobic response in gnd and sdhCB that involved a complex interaction of TFs ArcA, CRP, SoxR, SoxS, Fur, and Fnr, and small RNAs fnrS and ryhB (Panels E-H). E) Regulatory schematic of the signal transduction cascade triggered by the oxidized status of the membrane bound quinones ubiquinone (q8 and q8h2) and menaquinols (mql8, mqn8, 2dmmql8, and 2dmmq8), and anaerobic metabolites (e.g., lac-D). F) Regulatory interaction diagram between the different regulators. Metabolite and expression profiles of key components involved in the regulatory cascade for gnd (Panel G) and sdhCB (Panel H). Note the similar downregulation of fnr in response to ArcA activation through increased levels of lac-L and changes in the oxidized status of the membrane bound quinones, the upregulation of fnrS in response to activation of CRP-cAMP through increased levels of cAMP, and the downregulation of soxS in uGnd and uSdhCB.

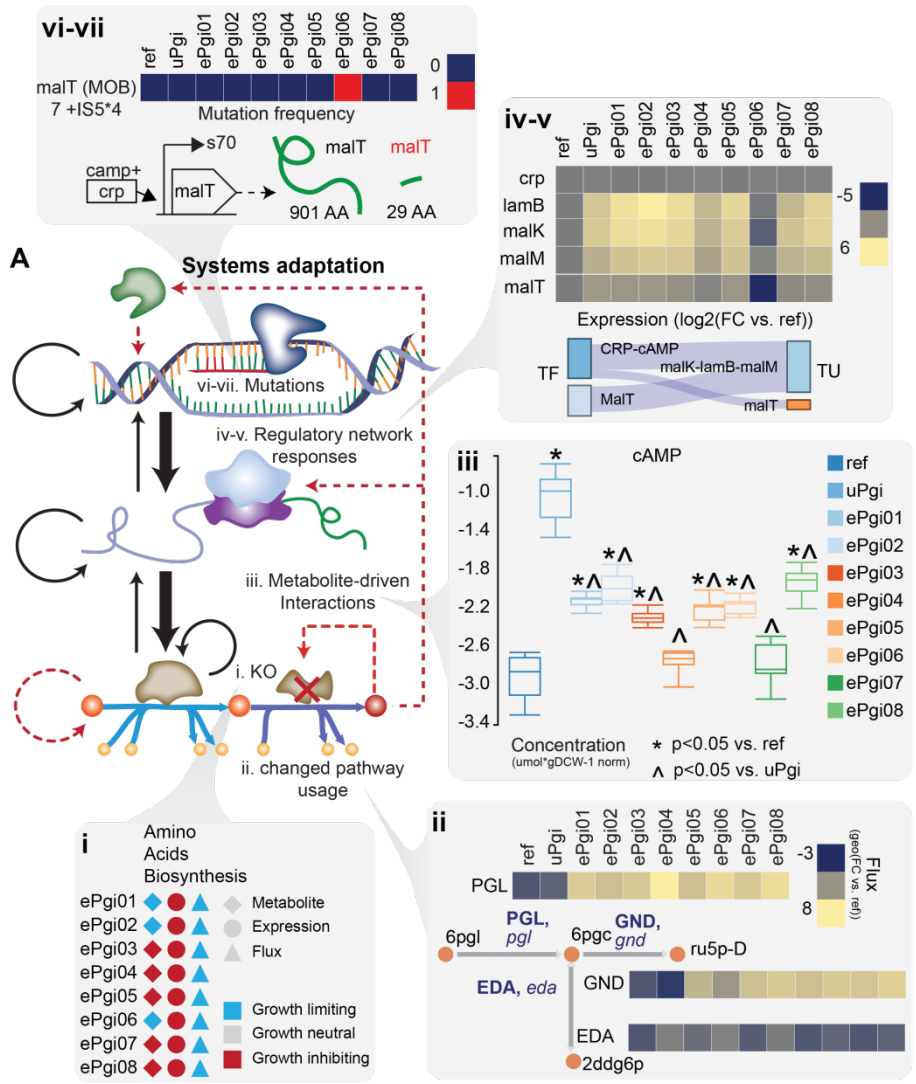

Figure 8.7: A model of systems adaptation and general principles of ALE that were revealed. A) A model of biological systems adaptation following the KO of key metabolic enzymes. The pgi lineages are used to exemplify each of the mechanisms (i-vii) found to occur during adaptive evolution. *i. Component imbalance retarded fitness:* Metabolites were found to be a mix of growth limiting and growth inhibiting, expression was found to be growth inhibiting, and fluxes were found to be growth limiting in amino acid biosynthesis pathways in the pgi lineages. *ii. Suboptimal pathway usage limited allocation of carbon to biomass precursors:* The forced flux through the oxidative pentose phosphate pathway in PGI was re-distributed through the oxidative and nonoxidative branches. *iii. Perturbed metabolite levels triggered transcription regulatory network responses in the uKOs:* cAMP levels were highly elevated in uPgi and restored to varying degrees in the ePGI strains. *iv. Transcription factor responses resulted in a misallocation of resources or amplification of processes reducing fitness in uKOs*: The elevation in cAMP levels resulted in the up-regulation of operons that are positively regulated by CRP in pgi, including the mal operons that encode enzymes for glycogen synthesis and turnover. *v. Alternate regulatory mechanisms abrogated counterproductive transcription factor responses: vi. Mutations selected during adaptive evolution re-wired many counterproductive system responses:* A mutation that truncated the *mal* operon activator, MalT, peptide was found in ePgi06 that appears to silence expression of the *mal* genes. *vii. Mutations selected during adaptive evolution also introduced innovations that targeted specific pathway or metabolite imbalances.*

Table 8.1: Counts of significantly different network components found for each evolved knockout relative to the unevolved knockout. Counts of metabolomic, transcriptomic, and fluxomic data are given as the average and standard deviation of the percent of significant features compared to all features measured for the lineage; counts for mutations are given as the average and standard deviation of the number of significant features (See Methods for criteria for significance).

	Metabolomics		**Transcriptomics**		**Fluxomics**		**Mutations**	
Comparison	**Ave % of total**	**STD % of total**	**Ave % of total**	**STD % of total**	**Ave % of total**	**STD % of total**	**Ave**	**STD**
ref vs. eRefs	2.2	0.0	0.4	0.1	0.0	0.0	6.5	0.7
uGnd vs eGnd	41.1	1.1	14.8	0.3	17.0	0.8	10.7	7.4
uPgi vs ePgi	29.9	5.2	3.8	2.8	43.0	3.8	14.0	7.3
uPtsHIcrr vs. ePtsHIcrr	20.8	2.8	19.1	2.7	38.3	1.9	12.0	4.8
uSdhCB vs. eSdhCB	24.4	5.9	10.4	0.5	13.8	5.5	7.7	2.3
uTpiA vs. eTpiA	36.9	2.9	10.5	0.8	40.2	0.1	19.7	7.4

Table 8.2: General principles of ALE uncovered and their implications

General principle	Implication
The dominant mode of adaptation to perturbation involved general and perturbation-specific network responses that were often re-balanced and modulated during the evolution process.	The initial network response to perturbation is sub-optimal, and requires adaptive evolution to re-optimize and achieve homeostasis.
The network state of each endpoint had unique and quantifiable differences. These differences were often attributed to mutations selected for during adaptation.	There is a diversity of system configurations that can be found during evolution to achieve the same physiological goal. While many omics measurements are restored, there is a subset that determines the uniqueness of alternate evolutionary outcomes.
System re-optimization during adaptation involved both proximal and distal changes relative to the location of perturbation (i.e., deleted reaction) in the network that reflected coordinated interaction of many layers of system components (i.e., metabolites, proteins, RNA, DNA, etc.). The primary drivers behind these changes were metabolites.	Biological systems should not only be analyzed from a component perspective, but must also be understood on a systems level where each component has multiple and often non-intuitive functions that extend beyond our current annotations.

References:

1. Alexa, A., and Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. R Package Version *2*.

2. Alvarez, A.F., and Georgellis, D. (2010). In vitro and in vivo analysis of the ArcB/A redox signaling pathway. Methods Enzymol. *471*, 205–228.

3. Applebee, M.K., Joyce, A.R., Conrad, T.M., Pettigrew, D.W., and Palsson, B.Ø. (2011). Functional and metabolic effects of adaptive glycerol kinase (GLPK) mutants in Escherichia coli. J. Biol. Chem. *286*, 23150–23159.

4. Balderas-Hernández, V.E., Sabido-Ramos, A., Silva, P., Cabrera-Valladares, N., Hernández-Chávez, G., Báez-Viveros, J.L., Martínez, A., Bolívar, F., and Gosset, G. (2009). Metabolic engineering for improving anthranilate synthesis from glucose in Escherichia coli. Microb. Cell Fact. *8*, 19.

5. Balsalobre, C., Johansson, J., and Uhlin, B.E. (2006). Cyclic AMP-dependent osmoregulation of crp gene expression in Escherichia coli. J. Bacteriol. *188*, 5935–5944.

6. Beauchene, N.A., Myers, K.S., Chung, D., Park, D.M., Weisnicht, A.M., Keleş, S., and Kiley, P.J. (2015). Impact of Anaerobiosis on Expression of the Iron-Responsive Fur and RyhB Regulons. MBio *6*, e01947–15.

7. Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

8. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

9. Bernard, T., Bridge, A., Morgat, A., Moretti, S., Xenarios, I., and Pagni, M. (2014). Reconciliation of metabolites and biochemical reactions for metabolic networks. Brief. Bioinform. *15*, 123–135.

10. Bobrovskyy, M., and Vanderpool, C.K. (2016). Diverse mechanisms of post-transcriptional repression by the small RNA regulator of glucose-phosphate stress. Mol. Microbiol. *99*, 254–273.

11. Bolten, C.J., Kiefer, P., Letisse, F., Portais, J.-C., and Wittmann, C. (2007). Sampling for metabolome analysis of microorganisms. Anal. Chem. *79*, 3843–3849.

12. Bordbar, A., Monk, J.M., King, Z.A., and Palsson, B.O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat.

Rev. Genet. *15*, 107–120.

13. Bradley, M.D., Beach, M.B., de Koning, A.P.J., Pratt, T.S., and Osuna, R. (2007). Effects of Fis on Escherichia coli gene expression during different growth stages. Microbiology *153*, 2922–2940.

14. Carlson, M. (2013). GO. db: A set of annotation maps describing the entire. Gene Ontology. 2013. R Package Version *3*.

15. Carroll, S.M., and Marx, C.J. (2013). Evolution after introduction of a novel metabolic pathway consistently leads to restoration of wild-type physiology. PLoS Genet. *9*, e1003427.

16. Charusanti, P., Conrad, T.M., Knight, E.M., Venkataraman, K., Fong, N.L., Xie, B., Gao, Y., and Palsson, B.Ø. (2010). Genetic basis of growth adaptation of Escherichia coli after deletion of pgi, a major metabolic gene. PLoS Genet. *6*, e1001186.

17. Chávez-Béjar, M.I., Báez-Viveros, J.L., Martínez, A., Bolívar, F., and Gosset, G. (2012/7). Biotechnological production of l-tyrosine and derived compounds. Process Biochem. *47*, 1017–1026.

18. Chen, Z., Lewis, K.A., Shultzaberger, R.K., Lyakhov, I.G., Zheng, M., Doan, B., Storz, G., and Schneider, T.D. (2007). Discovery of Fur binding site clusters in Escherichia coli by information theory models. Nucleic Acids Res. *35*, 6762–6777.

19. Cho, B.-K., Knight, E.M., Barrett, C.L., and Palsson, B.Ø. (2008). Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. Genome Res. *18*, 900–910.

20. Cho, B.-K., Federowicz, S., Park, Y.-S., Zengler, K., and Palsson, B.Ø. (2011a). Deciphering the transcriptional regulatory logic of amino acid metabolism. Nat. Chem. Biol. *8*, 65–71.

21. Cho, B.-K., Federowicz, S.A., Embree, M., Park, Y.-S., Kim, D., and Palsson, B.Ø. (2011b). The PurR regulon in Escherichia coli K-12 MG1655. Nucleic Acids Res. *39*, 6456–6464.

22. Cho, S., Cho, Y.-B., Kang, T.J., Kim, S.C., Palsson, B., and Cho, B.-K. (2015). The architecture of ArgR-DNA complexes at the genome-scale in Escherichia coli. Nucleic Acids Res. *43*, 3079–3088.

23. Conrad, T.M., Lewis, N.E., and Palsson, B.Ø. (2011). Microbial laboratory evolution in the era of genome-scale science. Mol. Syst. Biol. *7*, 509.

24. Cooper, T.F., Rozen, D.E., and Lenski, R.E. (2003). Parallel changes in

gene expression after 20,000 generations of evolution in Escherichia coli. Proceedings of the National Academy of Sciences *100*, 1072–1077.

25. Cooper, T.F., Remold, S.K., Lenski, R.E., and Schneider, D. (2008). Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli. PLoS Genet. *4*, e35.

26. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature *429*, 92–96.

27. Dahl, J.L., Wei, B.Y., and Kadner, R.J. (1997). Protein phosphorylation affects binding of the Escherichia coli transcription activator UhpA to the uhpT promoter. J. Biol. Chem. *272*, 1910–1919.

28. Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. *97*, 6640–6645.

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

30. Deutscher, J. (2008). The mechanisms of carbon catabolite repression in bacteria. Curr. Opin. Microbiol. *11*, 87–93.

31. Dragosits, M., and Mattanovich, D. (2013). Adaptive laboratory evolution -- principles and applications for biotechnology. Microb. Cell Fact. *12*, 64.

32. Federowicz, S., Kim, D., Ebrahim, A., Lerman, J., Nagarajan, H., Cho, B.-K., Zengler, K., and Palsson, B. (2014). Determining the control circuitry of redox metabolism at the genome-scale. PLoS Genet. *10*, e1004264.

33. 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., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

34. Flores, N., Xiao, J., Berry, A., Bolivar, F., and Valle, F. (1996). Pathway engineering for the production of aromatic compounds in Escherichia coli. Nat. Biotechnol. *14*, 620–623.

35. Fong, S.S., Joyce, A.R., and Palsson, B.Ø. (2005a). Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res. *15*, 1365–1372.

36. Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D., and Palsson, B.O. (2005b). In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol. Bioeng. *91*, 643–648.

37. Fong, S.S., Nanchen, A., Palsson, B.O., and Sauer, U. (2006). Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. J. Biol. Chem. *281*, 8024–8033.

38. Fuentes, L.G., Lara, A.R., Martínez, L.M., Ramírez, O.T., Martínez, A., Bolívar, F., and Gosset, G. (2013). Modification of glucose import capacity in Escherichia coli: physiologic consequences and utility for improving DNA vaccine production. Microb. Cell Fact. *12*, 42.

39. Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J.S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J.A., (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. *44*, D133–D143.

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

41. González Barrios, A.F., Zuo, R., Hashimoto, Y., Yang, L., Bentley, W.E., and Wood, T.K. (2006). Autoinducer 2 controls biofilm formation in Escherichia coli through a novel motility quorum-sensing regulator (MqsR, B3022). J. Bacteriol. *188*, 305–316.

42. Gray, M.J., Li, Y., Leichert, L.I.-O., Xu, Z., and Jakob, U. (2015). Does the Transcription Factor NemR Use a Regulatory Sulfenamide Bond to Sense Bleach? Antioxid. Redox Signal. *23*, 747–754.

43. Gresham, D., Desai, M.M., Tucker, C.M., Jenq, H.T., Pai, D.A., Ward, A., DeSevo, C.G., Botstein, D., and Dunham, M.J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. *4*, e1000303.

44. Grossmann, S., Bauer, S., Robinson, P.N., and Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. Bioinformatics *23*, 3024–3031.

45. Gunasekara, S.M., Hicks, M.N., Park, J., Brooks, C.L., Serate, J., Saunders, C.V., Grover, S.K., Goto, J.J., Lee, J.-W., and Youn, H. (2015).

Directed evolution of the Escherichia coli cAMP receptor protein at the cAMP pocket. J. Biol. Chem. *290*, 26587–26596.

46. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. *41*, D456–D463.

47. He, B., Shiau, A., Choi, K.Y., Zalkin, H., and Smith, J.M. (1990). Genes of the Escherichia coli pur regulon are negatively controlled by a repressor-operator interaction. J. Bacteriol. *172*, 4555–4562.

48. Hermsen, R., Okano, H., You, C., Werner, N., and Hwa, T. (2015). A growth-rate composition formula for the growth of E. coli on co-utilized carbon substrates. Mol. Syst. Biol. *11*, 801.

49. Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R., Albert, T.J., Blattner, F.R., van den Boom, D., Cantor, C.R., (2006). Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat. Genet. *38*, 1406–1412.

50. Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A Program for Missing Data. J. Stat. Softw. *45*, 1–47.

51. Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38, 27–28.

52. Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature *420*, 186–189.

53. Jackson, D.W., Simecka, J.W., and Romeo, T. (2002). Catabolite repression of Escherichia coli biofilm formation. J. Bacteriol. *184*, 3406–3410.

54. Janausch, I.G., Garcia-Moreno, I., Lehnen, D., Zeuner, Y., and Unden, G. (2004). Phosphorylation and DNA binding of the regulator DcuR of the fumarate-responsive two-component system DcuSR of Escherichia coli. Microbiology *150*, 877–883.

55. Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., (2013). EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. *41*, D605–D612.

56. Kim, D., Hong, J.S.-J., Qiu, Y., Nagarajan, H., Seo, J.-H., Cho, B.-K., Tsai,

S.-F., and Palsson, B.Ø. (2012). Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. PLoS Genet. *8*, e1002867.

57. Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J., and Smith, V.A. (2015). Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci. Rep. *5*, 10775.

58. Kreuzer, K.N. (2013). DNA damage responses in prokaryotes: regulating gene expression, modulating growth patterns, and manipulating replication forks. Cold Spring Harb. Perspect. Biol. *5*, a012674.

59. Kvitek, D.J., and Sherlock, G. (2011). Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet. *7*, e1002056.

60. LaCroix, R.A., Sandberg, T.E., O'Brien, E.J., Utrilla, J., Ebrahim, A., Guzman, G.I., Szubin, R., Palsson, B.O., and Feist, A.M. (2015). Use of Adaptive Laboratory Evolution To Discover Key Mutations Enabling Rapid Growth of Escherichia coli K-12 MG1655 on Glucose Minimal Medium. Appl. Environ. Microbiol. *81*, 17–30.

61. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Bowtie: an ultrafast memory-efficient short read aligner. Genome Biol. *10*, R25.

62. Larson, T.J., Cantwell, J.S., and van Loo-Bhattacharya, A.T. (1992). Interaction at a distance between multiple operators controls the adjacent, divergently transcribed glpTQ-glpACB operons of Escherichia coli K-12. J. Biol. Chem. *267*, 6114–6121.

63. Lavrrar, J.L., Christoffersen, C.A., and McIntosh, M.A. (2002). Fur-DNA interactions at the bidirectional fepDGC-entS promoter region in Escherichia coli. J. Mol. Biol. *322*, 983–995.

64. Lenski, R.E., Wiser, M.J., Ribeck, N., Blount, Z.D., Nahum, J.R., Morris, J.J., Zaman, L., Turner, C.B., Wade, B.D., Maddamsetti, R., (2015). Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with Escherichia coli. Proc. Biol. Sci. *282*, 20152292.

65. Link, H., Anselment, B., and Weuster-Botz, D. (2008). Leakage of adenylates during cold methanol/glycerol quenching of Escherichia coli. Metabolomics *4*, 240–247.

66. Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and

protein in complex biological samples. FEBS Lett. *583*, 3966–3973.

67. Majdalani, N., and Gottesman, S. (2005). The Rcs phosphorelay: a complex signal transduction system. Annu. Rev. Microbiol. *59*, 379–405.

68. Maloney, P.C., Ambudkar, S.V., Anatharam, V., Sonna, L.A., and Varadhachary, A. (1990). Anion-exchange mechanisms in bacteria. Microbiol. Rev. *54*, 1–17.

69. Martínez-Antonio, A., and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol. *6*, 482–489.

70. McCloskey, D., Gangoiti, J.A., King, Z.A., Naviaux, R.K., Barshop, B.A., Palsson, B.O., and Feist, A.M. (2014a). A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. Biotechnol. Bioeng. *111*, 803–815.

71. McCloskey, D., Utrilla, J., Naviaux, R.K., Palsson, B.O., and Feist, A.M. (2014b). Fast Swinnex filtration (FSF): a fast and robust sampling and extraction method suitable for metabolomics analysis of cultures grown in complex media. Metabolomics *11*, 198–209.

72. McCloskey, D., Gangoiti, J.A., Palsson, B.O., and Feist, A.M. (2015). A pH and solvent optimized reverse-phase ion-paring-LC–MS/MS method that leverages multiple scan-types for targeted absolute quantification of intracellular metabolites. Metabolomics *11*, 1338–1350.

73. McCloskey, D., Young, J.D., Xu, S., Palsson, B.O., and Feist, A.M. (2016a). MID Max: LC-MS/MS Method for Measuring the Precursor and Product Mass Isotopomer Distributions of Metabolic Intermediates and Cofactors for Metabolic Flux Analysis Applications. Anal. Chem. *88*, 1362–1370.

74. McCloskey, D., Young, J.D., Xu, S., Palsson, B.O., and Feist, A.M. (2016b). Modeling Method for Increased Precision and Scope of Directly Measurable Fluxes at a Genome-Scale. Anal. Chem. *88*, 3844–3852.

75. McDonald, M.J., Gehrig, S.M., Meintjes, P.L., Zhang, X.-X., and Rainey, P.B. (2009). Adaptive divergence in experimental populations of Pseudomonas fluorescens. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. Genetics *183*, 1041–1053.

76. Megchelenbrink, W., Huynen, M., and Marchiori, E. (2014). <italic>optGpSampler</italic>: An Improved Tool for Uniformly Sampling

the Solution-Space of Genome-Scale Metabolic Networks. PLoS One *9*, e86587.

77. Méhi, O., Bogos, B., Csörgő, B., Pál, F., Nyerges, A., Papp, B., and Pál, C. (2014). Perturbation of iron homeostasis promotes the evolution of antibiotic resistance. Mol. Biol. Evol. *31*, 2793–2804.

78. Meng, L.M., Kilstrup, M., and Nygaard, P. (1990). Autoregulation of PurR repressor synthesis and involvement of purR in the regulation of purB, purC, purL, purMN and guaBA expression in Escherichia coli. Eur. J. Biochem. *187*, 373–379.

79. Mevik, B.H., and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. J. Stat. Softw. *18*, 1–23.

80. Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. *44*, D523–D526.

81. Nelson, D.L., and Cox, M.M. (2013). Lehninger Principles of Biochemistry (W.H. Freeman).

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

83. Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., and Palsson, B.Ø. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. Mol. Syst. Biol. *7*, 535.

84. Ozyamak, E., de Almeida, C., de Moura, A.P.S., Miller, S., and Booth, I.R. (2013). Integrated stress response of Escherichia coli to methylglyoxal: transcriptional readthrough from the nemRA operon enhances protection through increased expression of glyoxalase I. Mol. Microbiol. *88*, 936–950.

85. Pál, C., Papp, B., and Pósfai, G. (2014). The dawn of evolutionary genome engineering. Nat. Rev. Genet. *15*, 504–512.

86. Plucain, J., Hindré, T., Le Gac, M., Tenaillon, O., Cruveiller, S., Médigue, C., Leiby, N., Harcombe, W.R., Marx, C.J., Lenski, R.E., (2014). Epistasis and Allele Specificity in the Emergence of a Stable Polymorphism in Escherichia coli. Science *343*, 1366–1369.

87. Richards, G.R., Patel, M.V., Lloyd, C.R., and Vanderpool, C.K. (2013). Depletion of glycolytic intermediates plays a key role in glucose-phosphate stress in Escherichia coli. J. Bacteriol. *195*, 4816–4825.

88. Rocke, D., Tillinghast, J., Durbin-Johnson, B., and Wu, S.L. LMGene Software for Data Transformation and Identification of Differentially Expressed Genes in Gene Expression Arrays. R package version 2.4. 0.

89. Sambrook, J., and Russell, D.W. (2001). Molecular cloning: a laboratory manual 3rd edition. Coldspring-Harbour Laboratory Press, UK.

90. Sandberg, T.E., Pedersen, M., LaCroix, R.A., Ebrahim, A., Bonde, M., Herrgard, M.J., Palsson, B.O., Sommer, M., and Feist, A.M. (2014). Evolution of Escherichia coli to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Mol. Biol. Evol. *31*, 2647–2662.

91. Schellenberger, J., and Palsson, B.Ø. (2009). Use of randomized sampling for analysis of metabolic networks. J. Biol. Chem. *284*, 5457–5461.

92. Seo, S.W., Kim, D., Szubin, R., and Palsson, B.O. (2015a). Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in Escherichia coli K-12 MG1655. Cell Rep. *12*, 1289–1299.

93. Seo, S.W., Kim, D., O'Brien, E.J., Szubin, R., and Palsson, B.O. (2015b). Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in Escherichia coli. Nat. Commun. *6*, 7970.

94. Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164–1167.

95. Sun, Y., and Vanderpool, C.K. (2013). Physiological consequences of multiple-target regulation by the small RNA SgrS in Escherichia coli. J. Bacteriol. *195*, 4804–4815.

96. Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D., and Gaut, B.S. (2012). The molecular diversity of adaptive convergence. Science *335*, 457–461.

97. Tenaillon, O., Barrick, J.E., Ribeck, N., Deatherage, D.E., Blanchard, J.L., Dasgupta, A., Wu, G.C., Wielgoss, S., Cruveiller, S., Médigue, C., (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. Nature *536*, 165–170.

98. Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D.L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically

sustained drug selection. Nat. Genet. *44*, 101–105.

99. Torres-Cabassa, A.S., and Gottesman, S. (1987). Capsule synthesis in Escherichia coli K-12 is regulated by proteolysis. J. Bacteriol. *169*, 981–989.

100. Tramonti, A., De Canio, M., and De Biase, D. (2008). GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. Mol. Microbiol. *70*, 965–982.

101. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

102. Umezawa, Y., Shimada, T., Kori, A., Yamada, K., and Ishihama, A. (2008). The uncharacterized transcription factor YdhM is the regulator of the nemA gene, encoding N-ethylmaleimide reductase. J. Bacteriol. *190*, 5890–5897.

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

104. Valgepea, K., Adamberg, K., and Vilu, R. (2011). Decrease of energy spilling in Escherichia coli continuous cultures with rising specific growth rate and carbon wasting. BMC Syst. Biol. *5*, 106.

105. Vanderpool, C.K., and Gottesman, S. (2004). Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. Mol. Microbiol. *54*, 1076–1089.

106. Vanderpool, C.K., and Gottesman, S. (2007). The novel transcription factor SgrR coordinates the response to glucose-phosphate stress. J. Bacteriol. *189*, 2238–2248.

107. Vital-Lopez, F.G., Wallqvist, A., and Reifman, J. (2013). Bridging the gap between gene expression and metabolic phenotype via kinetic models. BMC Syst. Biol. *7*, 63.

108. Weickert, M.J., and Adhya, S. (1993). The galactose regulon of Escherichia coli. Mol. Microbiol. *10*, 245–251.

109. Weinstein-Fischer, D., and Altuvia, S. (2007). Differential regulation of Escherichia coli topoisomerase I by Fis. Mol. Microbiol. *63*, 1131–1144.

110. Weston, L.A., and Kadner, R.J. (1988). Role of uhp genes in expression of the Escherichia coli sugar-phosphate transport system. J. Bacteriol. *170*, 3375–3383.

111. Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. *5*, 17.

112. Xu, D., and Zhang, Y. (2013). Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. Sci. Rep. *3*, 1895.

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

114. Young, J.D. (2014). INCA: a computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics *30*, 1333–1335.

115. Zhu, G., Golding, G.B., and Dean, A.M. (2005). The selective cause of an ancient adaptation. Science *307*, 1279–1282.

CHAPTER 9:

Conclusion

Much progress has been made in establishing the causality of mutations that occur during adaptive laboratory evolution (ALE) on organism physiology. In contrast, little progress has been made in detailing the mechanisms and overarching principles of evolution that govern the adaptive process. Using *E. coli* as a model organism, a set of gene knockouts in key metabolic pathways, and ALE, this thesis sought to lay the ground work for identifying the underlying mechanisms and principles of adaptation. First, a workflow for measuring intracellular metabolite concentration for over 100 metabolites in central carbohydrate metabolism, amino acid metabolism, cofactor and nucleotide biosynthesis and energy metabolism from sampling and extraction to separation and acquisition is described. Second, a workflow for measuring over 74 unique intracellular mass isotopomer distributions (MIDs) for metabolic flux analysis (MFA) at the genome-scale is described. And finally, a platform that integrates the above metabolomics and fluxomics data and additional –omics data types, with statistical and biochemical modelling techniques was developed and applied to interrogate, analyze, and interpret the adaptive changes found in response to metabolic perturbation. Mechanisms of adaptation were found, and several overarching principles of evolution were uncovered.