© 2017

Min Kyung Kim

ALL RIGHTS RESERVED

INFERENCE OF METABOLIC FLUX DISTRIBUTIONS FROM TRANSCRIPTOMIC DATA

by

MIN KYUNG KIM

A dissertation submitted to the Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Desmond S. Lun

And approved by

Dr. Desmond S. Lun

Dr. Caroline Colijn

Dr. Kwangwon Lee

Dr. Jongmin Nam

Dr. Nir Yakoby

Dr. Grace Brannigan

Camden, New Jersey

May 2017

ABSTRACT OF THE DISSERTATION

Inference of metabolic flux distributions from transcriptomic data

by MIN KYUNG KIM

Dissertation Director:
Dr. Desmond S. Lun

Studying changes in the cellular metabolism is important to understand what a living cell does for survival in response to external or internal perturbations. Even though intracellular metabolic flux (i.e. reaction rate) distributions are desirable data to this end, it is challenging to directly quantify fluxes through methods such as metabolic flux analysis using stable isotope labeling.

Several computational methods thus have been developed to infer system-level and condition-specific intracellular metabolic flux distributions, which are difficult to measure, from transcriptomic data, which are far easier to obtain. While powerful in many settings, existing methods have several practical shortcomings, and it is unclear which method has the best accuracy in general due to limited validation against experimentally measured fluxes.

In this thesis, we describe two computational methods called E-Flux2 (E-Flux method combined with minimization of l^2 norm) and SPOT (Simplified Pearson cOrrelation with Transcriptomic data), to be employed when a suitable biological objective is available and unavailable, respectively. Our method overcomes shortcomings of existing methods and combines desirable characteristics including applicability to a wide range of experimental conditions, production of a unique solution, fast running time, and the availability of a user-friendly implementation (at http://most.ccib.rutgers.edu/).

Most importantly, the predictive accuracy of our method was validated using the largest experimental dataset compiled to date, consisting of 43 experimental conditions of transcriptome measurements coupled with corresponding central carbon metabolic intracellular flux measurements (19 in *Escherichia coli*, 9 in *Saccharomyces cerevisiae*, 8 in *Bacillus subtilis*, 3 in *Synechocystis* sp. PCC 6803, 2 in *Synechococcus* sp. PCC 7002, and 2 in H4IIE rat hepatoma cell line). Our method provided as good as or better predictions than a representative sample of competing methods including pFBA (parsimonious flux balance analysis), in terms of the average of correlation between predicted and measured fluxes and of overall stability in predictions, especially in unicellular heterotrophic microorganisms. This makes our methods useful even in the absence of measured flux rates that allow some existing methods such as pFBA to be employed.

The goal of developing these computational tools is to better understand complex biological systems. Not only do the methods we developed contribute to advancing previous work, they have helped to answer biological research questions as well. In several collaborative research, our methods were used to understand the lipid accumulation mechanism of nitrogen-stressed *Phaeodactylum tricornutum* cells, verify the predictive power of a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002, and examine the metabolic impacts of RpiRc, a potent repressor of microbial toxins in *Staphylococcus aureus*.

.

DEDICATION

To my parents

Tae Soo Kim and Mi Ryung Ma

For their unconditional love and support.

ACKNOWLEDGEMENT

This work could not have been done without supports of countless people.

My first acknowledgement and a huge thank-you unhesitatingly goes to my thesis advisor, Prof. Desmond S. Lun for his academic, financial, and moral support over the past five years. I am grateful for his patience and knowledge in guiding me to grow as an independent researcher. I also thank him for providing me with many opportunities to collaborate with other scientists of different academic disciplines, which was sometimes challenging at the beginning but was always fully inspiring and rewarding in the end. I appreciate Prof. Lun not only for his help and guidance, but also for being a great person who listened to and understood any challenges I went through outside the lab.

I would like to give special thanks to my dissertation committee members, Prof. Caroline Colijn, Prof. Kwangwon Lee, Prof. Jongmin Nam, and Prof. Nir Yakoby, who have graciously taken time out of their busy schedules. I sincerely appreciate their time, insightful feedbacks, thoughtful comments, guidance and support throughout the course of this research.

I thank all of my collaborators who have directly or indirectly contributed to various parts of this work: Dr. Orly Levitan, Joomi Kim, and Prof. Paul G. Falkowski at Rutgers University-New Brunswick (section 3.1); Xiao Qian, Anagha Krishnan, and Prof. Gerard C. Dismukes at Rutgers University-New Brunswick (section 3.2); Dr. Divya Balasubramanian, Prof. Bo Shopsin, and Prof. Victor Torres at New York University School of Medicine (section 3.3); Dr. Nick Fyson and Prof. Caroline Colijn at Imperial College London (section 2.2.3).

I would like to express my appreciation to people I met in the United States, many of whom came from different parts of the globe, which significantly helped me to broaden my perspective about the world. I especially want to thank my lab members, CCIB friends and colleagues, and the CCIB faculty and staff for the support, friendship and helpful discussions. All of you made being thousands of miles away from home much easier. With regards to this research project, I especially thank Anatoliy Lane and James Kelley for their contribution to section 2.3.6, Kuhn Ip for his helpful comments on early drafts of the manuscript in Chapter 2, and Slim Karkar for his valuable feedbacks on presentations.

I would also like to thank my previous academic advisors in South Korea. I feel fortunate that I began my research career with Prof. Moon Jung Song, Prof. Yong Sik Kim, and Prof. Cheol-Min Ghim, all of whom emphasized research integrity and commitment to meaningful research.

Lastly, I am extremely grateful for my parents, Tae Soo Kim and Mi Ryung Ma, who have encouraged me to deliberate about how to make the world a better place through what I work on. I also thank my brother, Se Hwan Kim for encouraging conversations over video calls. None of this would have been possible without the support of my family.

This research was supported by the Samsung Advanced Institute of Technology (SAIT) through the Samsung Global Research Outreach (GRO) program and the National Science Foundation under grant number MCB-1515511.

NOTES

Chapter 1 is a modified version of material in **MK Kim and DS Lun (2014), Methods for integration of transcriptomic data in genome-scale metabolic models, Computational and structural biotechnology journal 11 (18), 59-65**. I was the primary author, while the corresponding author participated in and supervised the drafting of this review.

Chapter 2, in part, is a reprint of the material in **MK Kim, A Lane, JJ Kelley, DS Lun (2016), E-Flux2 and SPOT: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data, PLoS One 11 (6), e0157101**. I was the primary author, while the coauthors participated in the research that served as the basis for this study.

Chapter 3 integrates material from three publications in all of which I was a coauthor:

O Levitan, J Dinamarca, E Zelzion, DS Lun, LT Guerra, MK Kim, J Kim, BAS Van Mooy, D Bhattacharya, PG Falkowski (2015), Remodeling of intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress, PNAS 112 (2), 412-417;

X Qian, MK Kim, GK Kumaraswamy, A Agarwal, DS Lun, GC Dismukes (2016), Flux balance analysis of photoautotrophic metabolism: Uncovering new biological details of subsystems involved in cyanobacterial photosynthesis, BBA-Bioenergetics, 1858 (4) 276–287;

D Balasubramanian, EA Ohneck, J Chapman, A Weiss, MK Kim, T Reyes-Robles, J Zhong, LN Shaw, DS Lun, B Ueberheide, B Shopsin, VJ Torres (2016), *Staphylococcus aureus* coordinates leukocidin expression and pathogenesis by sensing metabolic fluxes via RpiRc, MBio 7 (3), e00818-16.

Chapter 4, in part, is based on the material in **MK Kim and DS Lun, Assessment of methods for inferring metabolic flux distributions from transcriptomic data in cells grown on different substrates, *In preparation***. I was the primary author and the corresponding author provided support in the research that served as the basis for this study.

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION ii

DEDICATION iv

ACKNOWLEDGEMENT v

NOTES vii

TABLE OF CONTENTS ix

CHAPTER 1: Introduction 1

1.1 ^{13}C-Metabolic Flux Analysis 1

1.2 Flux Balance Analysis for predicting system-level metabolic flux distributions 2

1.3 Integration of transcriptomic data for predicting condition-specific metabolic flux distributions 4

1.4 Summary of existing methods for inferring metabolic flux distributions from transcriptomic data based on four grouping criteria 8

1.4.1 Grouping criterion 1: Requirement for multiple gene expression datasets as input 9

1.4.2 Grouping criterion 2: Requirement for a threshold to define a gene's high/low expression 13

1.4.3 Grouping criterion 3: Requirement for *a priori* assumption of an appropriate objective function 16

1.4.4 Grouping criterion 4: Validation of predicted fluxes directly against measured intracellular fluxes 17

1.4.5 Summary and outlook 18

1.5 Objective of this study 20

CHAPTER 2: Development of E-Flux2 and SPOT 22

2.1 Background 22

2.2 Materials and Methods 27

2.2.1 Transcriptomic data, fluxomic data, and metabolic models used for this study 28

2.2.2 Creation of template metabolic models depending on carbon source information 31

2.2.3 Two different optimization strategies depending on the availability of biomass objective 32

2.2.3.1 E-Flux2 33

2.2.3.2 SPOT 36

2.2.4 Validation of the predictive accuracy of the algorithm using the measured fluxes 41

2.2.5 Algorithm implementation of our methods 43

2.3 Results 44

2.3.1 Validation of the accuracy of our predictions against measured intracellular fluxes 44

2.3.2 Comparison of correlation with competing methods 47

2.3.3 Detailed quantitative features of the predicted fluxes 50

2.3.4 Test of our methods on previous models of *E. coli* and *S. cerevisiae* 51

2.3.5 Measurement of the speed of our methods 56

2.3.6 Implementation of our methods in a user-friendly interface 57

2.4 Discussion 58

CHAPTER 3: Biological applications of our computational methods 60

3.1 Application 1: Understanding the lipid accumulation mechanism of *P. tricornutum* under nitrogen starvation 61

3.1.1 Background 61

3.1.2 Materials and Methods 62

3.1.2.1 Cultivation and experimental planning 62

3.1.2.2 Genome-scale metabolic network of *P. tricornutum* 63

3.1.2.3 Transcriptome measurements by RNA-Seq 64

3.1.2.4 Metabolic flux prediction 65

3.1.3 Results 66

3.1.4 Discussion 70

3.2 Application 2: Validation of the predictive power of a genome-scale metabolic model of *Synechococcus* sp. PCC 7002 73

3.2.1 Background 73

3.2.2 Materials and Methods 75

3.2.2.1 Metabolic network reconstruction 75

3.2.2.2 Transcriptome-constrained carbon metabolic fluxes under nitrogen -deprivation . 77

3.2.3 Results 77

3.2.4 Discussion 81

3.3 Application 3: Examination of metabolic roles of RpiRc, a potent repressor of leukocidins, in *S. aureus* 83

3.3.1 Background 83

3.3.2 Materials and Methods 84

3.3.2.1 Culture conditions and bacterial strains 84

3.3.2.2 RNA isolation, RNA-Seq, and data analyses 85

3.3.2.3 Computational metabolic flux prediction 86

3.3.3 Results 86

3.3.4 Discussion 94

CHAPTER 4: Assessment of our methods in cells grown on different substrates 96

4. 1 Background 96

4. 2 Materials and methods 98

4.2.1 Gene expression and flux dataset, and metabolic models used for this study 98

4.2.1.1 Data and model for *E. coli* 98

4.2.1.2 Data and model for *B. subtilis* 98

4.2.1.3 Data and model for *Synechocystis* sp. PCC 6803 99

4.2.1.4 Data and model for *Synechococcus* sp. PCC 7002 99

4.2.1.5 Data and model for H4IIE rat hepatoma cell 99

4.2.2 Computational metabolic flux prediction 100

4.2.3 Validation of the predictive accuracy of the methods used in this study 100

4. 3 Results 102

4. 4 Discussion 110

CHAPTER 5: Conclusion 114

Appendix 118

Appendix 1. A mathematical justification for dropping the $\boldsymbol{v}$ term in the objective function of SPOT 118

Appendix 2. A mathematical proof of the uniqueness of SPOT solutions 120

Appendix 3. Calculation of the possible range of correlation between the measured fluxes and the predicted fluxes 121

Bibliography 123

Curriculum Vitae 140

CHAPTER 1: Introduction

Intracellular metabolic reactions provide a cell with basic biochemical building blocks, energy, and a thermodynamically favorable environment to sustain its life. Because of the large connectivity inherent to metabolic networks via metabolites participating in multiple metabolic reactions, determination of system-level changes in intracellular metabolic fluxes of organisms is important for understanding the fundamental mechanisms of their metabolic responses to environmental or genetic perturbations [1,2].

1.1 ^{13}C-Metabolic Flux Analysis

^{13}C metabolic flux analysis (^{13}C-MFA) allows intracellular fluxes to be quantified experimentally. In this approach, cells are grown on ^{13}C-labeled substrates until the cells are at both metabolic steady state (i.e. when concentrations of metabolites remain stable over time) and isotopic steady state (i.e. when the isotope label is distributed throughout the network, and all isotopomer fractions are constant over time). Then the level of ^{13}C enrichment in metabolites of the cells is measured by mass spectrometry (MS) or nuclear magnetic resonance (NMR). Intracellular flux distribution is reconstituted from the ^{13}C enrichment patterns [3–8]. System-wide quantification of intracellular metabolic fluxes using ^{13}C-MFA, however, is challenging not only because of the extensive instrumentation required but also because of the limited number of fluxes and conditions that can be experimentally measured. Typically, ^{13}C-MFA focuses on central carbon metabolism [7–10].

1.2 Flux Balance Analysis for predicting system-level metabolic flux distributions

An alternative method that is widely used for system-level studies of metabolism is a computational modeling approach called flux balance analysis (FBA). FBA predicts metabolic flux distributions at steady state by making use of *in silico* genome-scale metabolic models [11]. These genome-scale metabolic models are assembled and manually-curated from annotated genome, biochemical, genetic, and cell phenotype data [11–13]. To use FBA, a genome-scale metabolic model is converted into a $m \times n$ stoichiometric matrix, S, where the rows in S correspond to the m metabolites of the metabolic network, and the columns represent the n reactions (Figure 1.1a). Each matrix element s_{ij}, indicates a stoichiometric coefficient, that is, the number of molecules of the i^{th} metabolite participating in the j^{th} reaction. $s_{ij} = 0$ means that the i^{th} metabolite is not involved, and a positive or a negative s_{ij} indicates that the i^{th} metabolite is a product or a reactant of the j^{th} reaction, respectively. Under the steady state assumption, the metabolic flux distribution can be represented mathematically by $S \cdot v = 0$, where v is a column vector whose elements are the unknown reaction rates (fluxes) through each of the reactions of S (Figure 1.1b). Since genome-scale metabolic models include all possible metabolic reactions implied by the genome annotation regardless of whether the annotated metabolic genes are expressed in a given environment, the resulting system $S \cdot v = 0$, is in general underdetermined [14,15].Thus, physiologically meaningful flux solutions need to be narrowed down from all the possible flux distributions by imposing additional constraints on the system and by optimizing certain objective functions when performing FBA (Figure 1.1c) [16]. The standard FBA involves solving the following linear optimization problem:

$$\max \; f'v$$
$$\text{subject to} \begin{cases} Sv = 0 \\ lb \le v \le ub \end{cases} \qquad (1)$$

where v is a flux vector representing the reaction rates of the n reactions in the network, f is a coefficient vector defining the organism's objective function, S is the stoichiometric matrix, and lb and ub are the minimum and maximum reaction rates through each reaction in v.

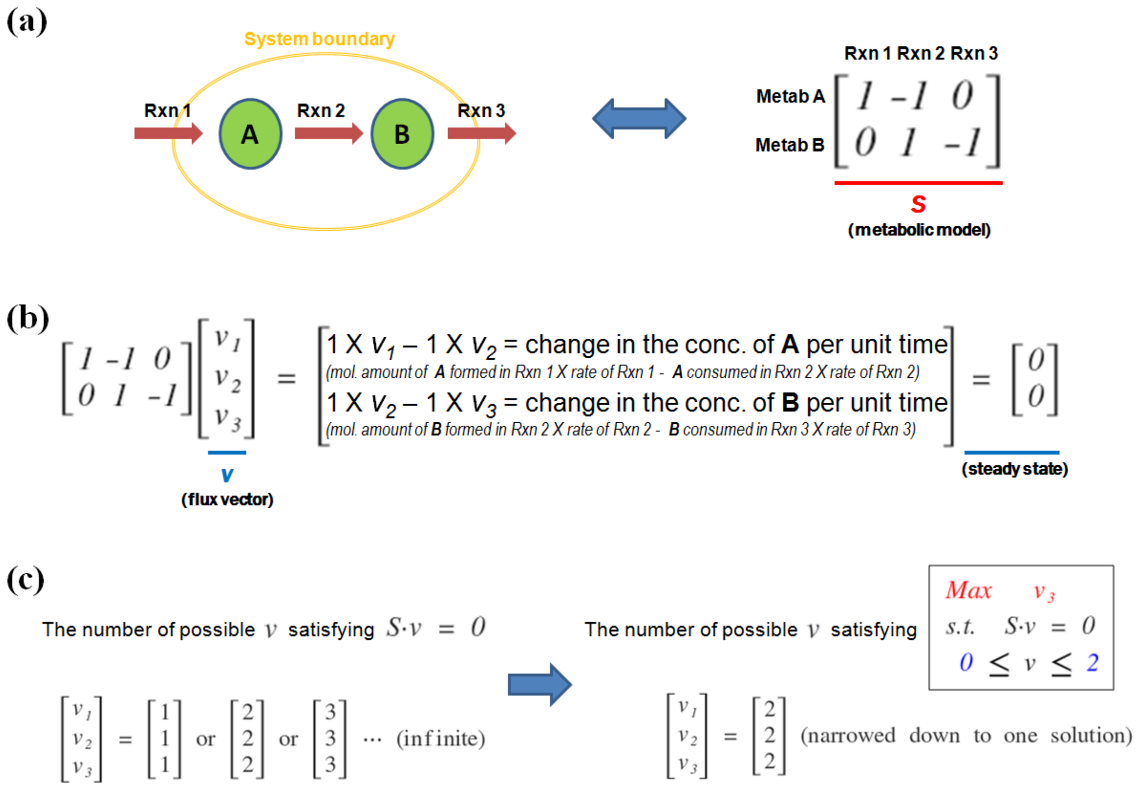

(a)
System boundary
Rxn 1
A
Rxn 2
B
Rxn 3
Rxn 1 Rxn 2 Rxn 3
Metab A
Metab B
1 -1 0
0 1 -1
S
(metabolic model)
(b)
1 -1 0
0 1 -1
v1 v2 v3
=
1 X v1 – 1 X v2 = change in the conc. of A per unit time
(mol. amount of A formed in Rxn 1 X rate of Rxn 1 - A consumed in Rxn 2 X rate of Rxn 2)
1 X v2 – 1 X v3 = change in the conc. of B per unit time
(mol. amount of B formed in Rxn 2 X rate of Rxn 2 - B consumed in Rxn 3 X rate of Rxn 3)
=
0 0
v
(flux vector)
(steady state)
(c)
The number of possible v satisfying S·v = 0
v1 v2 v3 = 1 1 1 or 2 2 2 or 3 3 3 ··· (infinite)
The number of possible v satisfying
Max v3
s.t. S·v = 0
0 ≤ v ≤ 2
v1 v2 v3 = 2 2 2 (narrowed down to one solution)

Fig 1.1. Flux Balance Analysis (FBA). Figure 1.1 illustrates how FBA works with an example of the simple network below consisting of two metabolites, A and B, and three metabolic reactions. (a) To use FBA, the network is converted into a stoichiometric matrix, $\boldsymbol{S}$, where the rows in $\boldsymbol{S}$ correspond to the metabolites of the metabolic network, and the columns represent the reactions. Each matrix element $\boldsymbol{s_{ij}}$, indicates a stoichiometric coefficient, that is, the number of molecules of the $\boldsymbol{i}^{\mathbf{th}}$ metabolite participating in the $\boldsymbol{j}^{\mathbf{th}}$ reaction. $\boldsymbol{s_{ij} = 0}$ means that the $\boldsymbol{i}^{\mathbf{th}}$ metabolite is not involved, and a positive or a negative $\boldsymbol{s_{ij}}$ indicates that the $\boldsymbol{i}^{\mathbf{th}}$ metabolite is a product or a reactant of the $\boldsymbol{j}^{\mathbf{th}}$ reaction, respectively. (b) Under the steady state assumption, the metabolic flux distribution can be represented mathematically by $\boldsymbol{S \cdot v = 0}$, where $\boldsymbol{v}$ is a column vector whose elements are the unknown reaction rates (fluxes) through each of the reactions of $\boldsymbol{S}$.(c) Since the resulting system, $\boldsymbol{S \cdot v = 0}$, is usually underdetermined, physiologically meaningful flux solutions need to be narrowed down from all the possible flux distributions by imposing additional constraints on the system (e.g. $\boldsymbol{0 \leq v \leq 2}$ in the figure) and by optimizing certain objective functions (e.g. $\boldsymbol{Max\ v_3}$ in the figure).

1.3 Integration of transcriptomic data for predicting condition-specific metabolic flux distributions

If the complete regulatory structure of an organism were known, it would be possible to produce context-specific constraints by computing which cellular components may be expressed in a given condition. However, the regulatory structure is unknown even for the relatively simple and extensively-studied bacterium, *E. coli*, partly due to the lack of comprehensive transcription unit information and because of the lack of information on the relationship between genotype and phenotype [17].

Recent advances in omics technologies have enabled quantitative monitoring of the abundance of biological molecules at various levels in a high-throughput manner [18] (Fig 1.2). In the absence of complete information on regulatory rules, omics data can be integrated with genome-scale metabolic models to improve their predictive power [19,20]. For this purpose, transcriptomic data, i.e. genome-wide mRNA expression profiling data, is useful in some points compared to other omics platforms. Fluxomics (i.e. ^{13}C-MFA) is the most direct measurement of metabolic phenotype, but has the disadvantages in that it is difficult to make measurements and only a limited number of fluxes that can be determined as mentioned above. Metabolomics can also be useful, but typically fluxes are more informative than metabolite concentrations themselves, and it is challenging to determine fluxes from metabolite concentrations since the relation between the amount of metabolites and metabolic fluxes is less straightforward compared to the other omic data. This is partly because each metabolite often participates in multiple metabolic reactions (e.g. ATP) and because the high concentration of a certain metabolite can be interpreted in more than two ways (i.e. high flux coming in towards, or low flux going out from that metabolite). Additionally, similar to fluxes, specific classes of metabolites such as lipids or labile chemicals easily metabolized are still demanding to measure [21,22]. Unlike first two omics data that cover a small share of all reactions in a genome-scale model, transcriptomics and proteomics are the platforms where a quantitative snapshot of molecular species at system-level is currently possible [23]. In addition, compared to metabolomic data, the relation between fluxes and transcriptome or proteome measurements is more straightforward: Even though the abundance of mRNA or protein of a certain metabolic enzyme does not always guarantee a high reaction rate of the corresponding metabolic reaction, a low level of them cannot lead to a high metabolic flux. However, proteomics is a relatively immature technology compared to transcriptomics. The accuracy with which protein concentrations can be determined is much lower than that with which mRNA concentrations can be determined. On the other hand, RNA amount changes can be precisely measured in a highly automated process at low cost in comparison with the amount of data gathered [24,25]. By integrating transcriptomics data with genome-scale metabolic models, we can potentially determine metabolic fluxes through a relatively simple and low-cost omics technology. If other omics technology especially proteomics technology becomes as mature (e.g. wide coverage at lower cost with less effort) as that of transcroptomics, most of the methods introduced in this paper could be applied to other omics data, too.

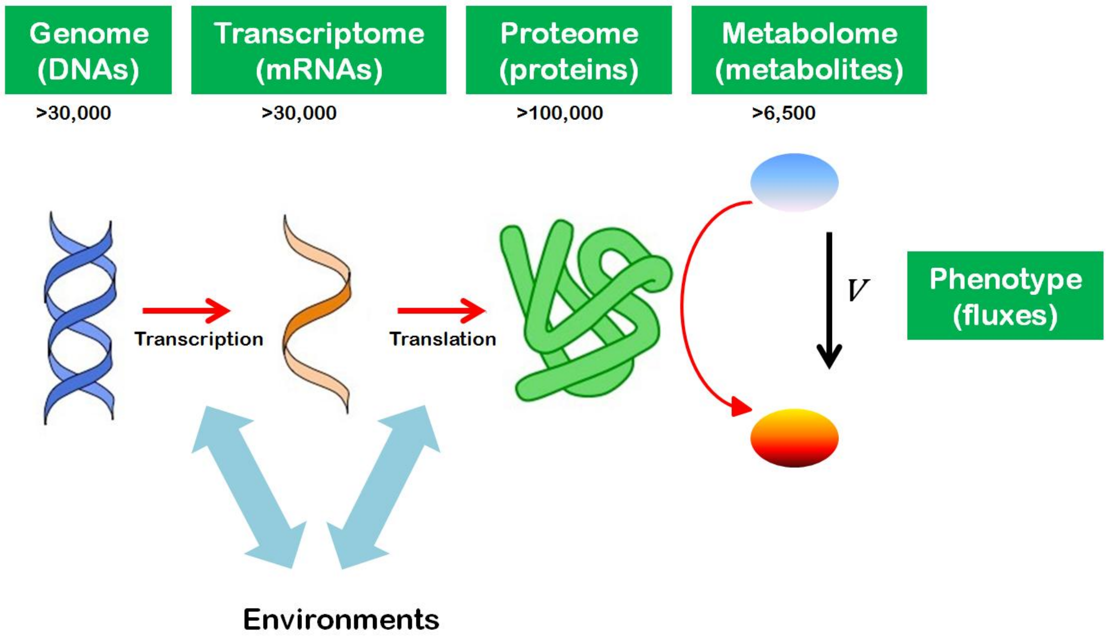


Fig 1.2. Functional levels of omic data in a biological system. This figure shows how genetic information flows from relatively static DNA sequences to dynamic phenotypes (fluxes, in this case) as a cell interacts with its environments. The numbers below each functional level shows the approximate quantity in human.

Not only do genome-scale models benefit from transcriptomic data in creating condition- and tissue- specific models, but transcriptomic data itself can also benefit by being integrated onto the models. Although a large amount of transcriptomic data is continuously being generated, gaining meaningful insight into the functioning of cellular processes from mRNA levels is challenging because of the functional layers in between the two, such as translation, post-translational modifications, mRNA/protein degradation, and enzyme activity regulation by effectors (inhibitors or activators) [14,23,26]. Genome-scale metabolic models are well-suited to inferring metabolic phenotype from genotype using transcriptomic data, since the models are comprehensive repositories of biochemical data for organisms that enable the description of gene-protein-reaction relationships [13,19]. Whereas correlations between mRNA and fluxes have been often found to be poor, approaches taking into account for the large connectivity of metabolites inherent to metabolic networks have been successful in linking gene expression level to metabolites [2,27–29]. This implies that the consideration of the metabolic network is essential to draw a predictive relation from transcript abundances to fluxes [23].

1.4 Summary of existing methods for inferring metabolic flux distributions from transcriptomic data based on four grouping criteria

For these reasons mentioned above, there have been previous studies to integrate transcriptomic data with genome-scale metabolic models, and some of these methods have been covered in recent reviews [12,14,18,30–32]. However, most of these reviews broadly introduce methods inferring metabolic fluxes from various kinds of omics data and are not focused specifically on transcriptomic data. In addition, some of the reviews do not include the most recent methods since transcriptomic data-driven metabolic modeling methods are being developed at a fast pace [32]. In this section, we focus on introducing methods for integrating transcriptomic data in genome-scale metabolic models, and we give a brief description of each one published to date. We exclude methods that require multi-omics datasets as input for an analysis even if they use transcriptomic data, because multi-omics studies are not common [33–36]. We categorize all methods that are covered in this section based on four different grouping criteria, and we evaluate which group of methods is more suitable from a practical perspective. Lastly, we discuss several limitations of existing methods that new methods need to overcome.

1.4.1 Grouping criterion 1: Requirement for multiple gene expression datasets as input

As the first criterion, methods for estimating metabolic flux from transcriptomic data can be grouped by how many gene expression datasets are required as input. There are two representative methods that need multiple transcriptomic datasets measured under two or more conditions for an analysis.

First, Probabilistic Regulation Of Metabolism (PROM) published in 2010 is a method that integrates regulatory and metabolic networks [37]. It calculates the probability of a metabolic target gene being expressed relative to the activity of its regulating transcription factor from a large dataset of gene expression data, and the flux maxima of the metabolic reaction associated with the metabolic target gene is constrained by a factor of this probability (Fig 1.3a). It has several advantages such as its ability to account for the presence of noise in the data, and to differentiate between a strong transcriptional regulator and a weak one. However, this method requires a large number of experimental datasets to calculate the probability of regulatory interactions between transcription factors and their target genes. It also requires *a priori* knowledge on transcription factor-target gene pairs. In the original paper, around 1,300 microarrays and 2,000 transcription factor-target interactions were used for *E. coli* and *M. tuberculosis*.

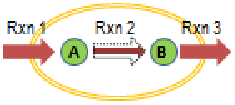
Second, Metabolic Adjustment by Differential Expression (MADE) published in 2011, was developed to overcome the issue of selecting a subjective user-supplied threshold in defining a gene's high and low expression states [38]. MADE creates a sequence of binary expression states using several datasets for differential gene expression so as to find the model that most closely reproduces the observed expression changes (Fig 1.3b). The principle of this method is that if the activity of a gene drastically changes from one condition to the other, the flux through the reaction controlled by that gene will change accordingly [39]. Using this method, the authors examined the metabolic effects of the transition from glucose- to glycerol-based growth in *S. cerevisiae* over the course of time. They showed that the binary expression state changes calculated by MADE matched 98.7% of the feasible observed gene expression transitions (83.5% of all expression transitions). They also showed that, accompanied by these expression state changes, the flux variability of the model was increased after the shift to glycerol.

The other methods described below use a single gene expression dataset for each experimental condition. One of the possible concerns of using a single transcriptomic dataset may be the lack of proportionality between transcript and flux levels. Accounting for relative gene expression changes from multiple datasets as an indicator of the flux reconfiguration might seem to provide a more meaningful description. However, a recent research paper shows that the methods that use relative expression levels does not necessarily give more accurate flux predictions [32]. Although both methods have advantages, the requirement for multiple sets of input data such as transcription regulatory information or different gene expression datasets to perform the analysis is more onerous from a practical point of view.

(a) PROM

	Set 1	Set 2	Set 3
Rxn 2's gene expr. Lv	0.88	0.23	0.12
Binarized data when Threshold = 0.5	1 (above 0.5)	0 (below 0.5)	0

****Probability** of Rxn 2 gene being expressed assuming its regulating transcription factor is active: (1+0+0)/(1+1+1) = 0.33



V_{max} through Rxn 2 = 0.33 X V_{max}

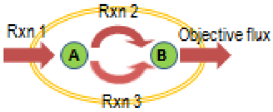
(b) MADE

*Observed gene expression level

	Set 1	Set 1 -> 2	**Set 2**	Set 2 -> 3	**Set 3**
Rxn 1	**223**	Decreased	**158**	Constant	**162**
Rxn 2	**174**	Decreased	**52**	Constant	**48**
Rxn 3	**23**	Increased	**88**	Increased	**102**

*MADE binary approximation

	Set 1	Set 2	Set 3
Rxn 1	1	1	1
Rxn 2	1	0	0
Rxn 3	0	1	1



*Although its gene expression level has been statistically significantly decreased (set 1 to set 2), **Rxn 1** (Input) should be always active/on (binary state = 1) for a functional model (viable objective flux) across all conditions.

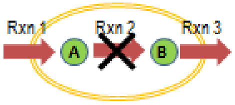
*Although its gene expression level has been statistically significantly increased (set 2 to set 3), **Rxn 3** is assigned as 1 since only a binary approximation is allowed.

*The resulting model indicates that A is transformed to B **by Rxn2** when **Set 1** was measured, and the flux is redirected **through Rxn 3** when **Set 2 & 3** were measured.

(c) Åkesson et al.

	Set 1
Rxn 2's gene expr. Lv	0

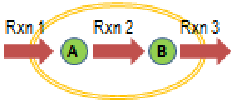
*A probe set for a gene is considered absent if it is undetected in replicates from independent cultures of the same condition.



(d) GIMME

	Set 1
Rxn 2's gene expr. Lv	11.4

Step 1. FBA



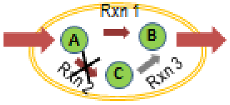
Let's say,
optimal flux of Rxn2 = 7

*If threshold = 11,
IS(Inconsistency Score) = 0
Because **expression > threshold**
*If threshold = 12,
IS= (optimal flux)*(threshold - data)
= 7*(12 - 11.4) = 4.2
*If threshold = 15,
IS= 7*(15 - 11.4) = 25.2
*GIMME finds a metabolic flux distribution whose ∑IS becomes **minimum**.

(e) iMAT

***low cutoff = 0.3** ***high cutoff = 0.7**	Set 1
Rxn 1's gene expr. Lv	**0.6**
Rxn 2's gene expr. Lv	0.2 (below low cutoff)
Rxn 3's gene expr. Lv	0.8 (above high cutoff)

*This method finds a metabolic flux distribution the most consistent with the gene expression data by maximizing the number of reactions highly-expressed and minimizing the number of reactions lowly-expressed.



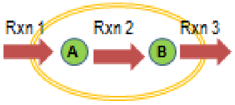
*Flux through Rxn 3 (high gene expr. But no flux) is considered to be **post-transcriptionally down-regulated**

(f) E-Flux

	Set 1
Rxn 2's gene expr. Lv	**20**

*Gene expression level determines flux limits of an **arbitrary unit**

If Rxn2 is irreversible,
$0 \le$ Rxn 2 flux $\le$ **20**
If Rxn2 is reversible,
$-20 \le$ Rxn 2 flux $\le$ **20**



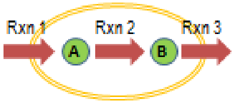
V_{max} through Rxn 2 = 20

(g) Dave Lee et al.

	Set 1
Rxn 2's gene expr. Lv	**20**

*This method predicts intracellular metabolic fluxes by minimizing the sum of absolute differences between fluxes and corresponding gene expression data.

e.g. (right figure)
when Rxn 2 flux = 20,
| Rxn 2 flux – 20| = 0 (minimum)



Rxn 2 flux = 20

Fig 1.3. Representative methods currently available for integration of transcriptomic data in genome-scale metabolic models. Figures (a) - (g) show how each method integrates gene expression data onto the models. (a) PROM binarizes the gene expression data according to a user-supplied threshold. Then, it calculates the probability of a metabolic target gene being expressed relative to the activity of its regulating transcription factor from a large dataset of gene expression data. The flux maxima of the metabolic reaction associated with the metabolic target gene is constrained by a factor of this probability. (b) MADE creates a sequence of binary expression states using several datasets for differential gene expression so as to find the model that most closely reproduces the observed expression changes.(c) Å kesson's method is one of the earliest methods to integrate genome-wide expression data into genome-scale metabolic models. In this method, the fluxes of reactions whose corresponding genes are not expressed are constrained as zero. (d) GIMME consists of a two-step procedure. First, the method finds a flux distribution that optimizes a given biological objective such as growth and/or ATP production using FBA. Then, the method minimizes the utilization of 'inactive' reactions whose corresponding mRNA transcript levels are below a given threshold.(e) iMAT discretized gene expression data into tri-valued expression states, representing either low, moderate or high expression in the condition studied according to a user-specified threshold. Then, the method finds an optimal metabolic flux distribution that is the most consistent with the discrete gene expression data by maximizing the number of flux-carrying reactions associated with highly expressed enzymes and minimizing the number of flux-carrying reactions that correspond to lowly-expressed enzymes. (f) E-Flux maps continuous gene expression levels into flux bound constraints according to gene-protein-reaction (GPR) associations. It uses transcriptomic data to set upper and lower bounds on metabolic fluxes so that reactions associated with more highly expressed genes will be allowed to have higher absolute flux values. (g) Dave Lee's method uses

transcriptomic data in the objective function. This method predicts intracellular metabolic fluxes by minimizing the deviation between the flux distribution and the transcriptomic data. The deviation was calculated by the sum of absolute differences between fluxes and corresponding gene expression data.

1.4.2 Grouping criterion 2: Requirement for a threshold to define a gene's high/low expression

As the second criterion, methods can be grouped by whether they use a user-supplied threshold. Some methods require discretization (e.g. -1, 0, 1), binarization (e.g. 1, 0), or classification (e.g. below/above threshold) of gene expression measurement data according to user-defined arbitrary thresholds to distinguish active and inactive states of the corresponding reactions. In addition to PROM, which is mentioned in the previous section, the following three methods also require thresholds.

An approach suggested by Åkesson *et al*. in 2004 is one of the earliest methods to integrate genome-wide expression data into genome-scale metabolic models [40]. In this method, the fluxes of reactions whose corresponding genes are not expressed are constrained as zero (Fig 1.3c). A probe set for a gene is considered absent if it is undetected in all three replicates from independent cultures of the same condition. Using this principle, they combined microarray measurements of gene expression from chemostat and batch cultivations of *S. cerevisiae* with a genome-scale model for yeast, *i*FF708 [41]. The computed metabolic flux distributions were compared to experimental values from ^{13}C-labeling experiments. The integration of expression data resulted in improved predictions of metabolic behavior in batch culture. Due to the Boolean nature of this method, failure in correctly detecting presence of lowly expressed genes may give rise to erroneous predictions.

Gene Inactivity Moderated by Metabolism and Expression (GIMME) introduced in 2008, creates a context-specific metabolic model that predicts the subset of reactions a cell is likely to use under particular conditions using gene expression data [42]. This method consists of a two-step procedure (Fig 1.3d). First, the method finds a flux distribution that optimizes a given biological objective such as growth and/or ATP production using FBA. Then, the method minimizes the utilization of 'inactive' reactions whose corresponding mRNA transcript levels are below a given threshold. By avoiding the use of below-threshold reactions that are inconsistent with the flux distribution of the first step, the method was used to find context-specific metabolic flux distributions that best fit physiological data in *E.coli* and human skeletal muscle cells.

The integrative Metabolic Analysis Tool (iMAT) implements a method proposed by Shlomi *et al.* in 2008, which was developed for tissue-specific modeling of metabolism in mammalian cells [43,44]. In this method, gene expression data is discretized into tri-valued expression states, representing either low, moderate or high expression in the condition studied according to a user-specified threshold (Fig 1.3e). Then, iMAT finds an optimal metabolic flux distribution that is the most consistent with the discrete gene expression data by maximizing the number of flux-carrying reactions associated with highly expressed enzymes and minimizing the number of flux-carrying reactions that correspond to lowly-expressed enzymes. This method does not require information on biomass composition or metabolite exchange. By integrating transcriptomic data with a global human metabolic model using this method, they predicted tissue-specific metabolic activity in ten different tissues. A method called EXAMO (EXploration of Alternative Metabolic Optima) is an extended version of iMAT that builds a context-specific model [45].

Tailored gene expression using user-defined thresholds may avoid data normalization issues [32]. However, using arbitrary thresholds may lead to subjective results that loses the fine-grained information for individual genes. This is because the specific threshold above which the level of gene expression indicates physiological activeness of corresponding reactions may vary across genes, conditions, or organisms. The following two methods incorporate continuous gene expression values without using thresholds.

E-Flux (as a combination of flux and expression) published in 2009 is a method that maps continuous gene expression levels into flux bound constraints according to gene-protein-reaction (GPR) associations [46,47]. It uses transcriptomic data to set upper and lower bounds on metabolic fluxes so that reactions associated with more highly expressed genes will be allowed to have higher absolute flux values (Fig 1.3f). The rationale behind E-flux is that, given a limited translational efficiency and a limited accumulation of enzyme over the time, the level of mRNA can be used as an approximate upper bound on the maximum amount of metabolic enzymes, and hence as a bound on reaction rates. Using this method, the authors correctly predicted decreased mycolic acid synthesis by seven of the eight known fatty acids inhibitors in *M. tuberculosis*. In a follow up study [47], they identified preferred carbon sources of *E. coli* that are not influenced by expression derived constraints.

An approach suggested by Lee *et al.* uses transcriptomic data in the objective function [48]. This method predicts intracellular metabolic fluxes by minimizing the deviation between the flux distribution and the transcriptomic data (Fig 1.3g). The deviation was calculated by the sum of absolute differences between fluxes and corresponding gene expression data. The assumption behind this method is that enzymatic transcript concentrations and metabolic fluxes can be related to each other, albeit in a complex manner, since the existence of a transcript is necessary but not sufficient for the presence or activity of its corresponding enzyme [49]. They compared this method against FBA, GIMME, and iMAT, showing a better accuracy in predicting experimentally measured exometabolic flux for *S. cerevisiae* cultures under two growth conditions. FALCON (Flux Assignment with Least absolute deviation Convex Objectives and Normalization) is a recently published, related method with improvements in time efficiency [50].

1.4.3 Grouping criterion 3: Requirement for *a priori* assumption of an appropriate objective function

The third feature that can distinguish the methods is whether a method requires the *a priori* assumption of an appropriate biological objective function.

Except for the method of Lee *et al*. and iMAT, the other methods described here need *a priori* knowledge of an appropriate objective function of the system such as biomass production rate. The biomass flux (i.e. the growth rate) is the most widely used objective function for FBA optimization problems since it is commonly assumed that, under given resources, efficient growth of a certain microorganism compared to its competitors is beneficial for its survival from an evolutionary perspective [51,52]. Indeed, the assumption of biomass flux maximization in FBA has successfully predicted metabolic behavior of various organisms in a number of studies [53,54]. Nevertheless, biomass flux may be unsuitable as an objective function for some organisms such as microorganisms with variable biomass composition, pathogens in dormancy or in latent phase, or cells of a multi-cellular organism [55]. Thus, in practical applications, we sometimes need methods like the method of Lee *et al*. and iMAT whose objective functions can be universally applied to a variety of organisms in cases where knowledge of the biological objective function is uncertain.

1.4.4 Grouping criterion 4: Validation of predicted fluxes directly against measured intracellular fluxes

The last distinction among the methods is the utilization of measured intracellular fluxes for the purpose of validation. Basically, the output of the methods described here is predicted intracellular metabolic flux distribution. With the exception of the method of Åkesson et al., none of these methods, however, have tested their predictive accuracy against experimentally measured intracellular fluxes. Lee et al. did attempt to validate their predictions for the intracellular fluxes indirectly using exometabolomic data by measuring changes in the concentration of extracellular metabolites. Nevertheless, considering that detailed information on the underlying mechanisms of metabolic responses is not accessible from extracellular physiological data, it would be preferable to validate predictive accuracy using measured intracellular fluxes [56].

Table 1 summarizes the features of the presented methods with regard to the four grouping criteria described so far.

Method	Requirements for multiple transcriptomic datasets as input	Requirement for a threshold to define a gene's high/low expression state	Requirement for *a priori* assumption of an appropriate objective function	Validation of predicted fluxes directly against measured intracellular fluxes
E-Flux	No	No	Yes	No
Lee et al.	No	No	No	No
Åkesson et al.	No	Yes	Yes	Yes (4 fluxes were used for validation)
GIMME	No	Yes	Yes	No
iMAT	No	Yes	No	No
PROM	Yes	Yes	Yes	No
MADE	Yes	No	Yes	No

Table 1. Summary of the features of previous methods according to four grouping criteria described in this paper. Desirable features from a practical perspective are shaded in green.

1.4.5 Summary and outlook

Given its many advantages, the integration of transcriptomic data in a genome-scale model is a promising method for predicting system-level intracellular metabolic fluxes.

From a practical perspective, we suggest that an ideal method satisfy all of the following criteria: a method that needs a single gene expression dataset as input; that utilizes continuous gene expression values without using arbitrary thresholds; that can be used even when an appropriate objective function is unknown; and whose predictive accuracy is validated against measured intracellular fluxes data.

Yet none of the surveyed methods satisfies all of the practical conditions. Lee's method seems to be the most practical method among them in that it achieves three of the four criteria for a practically ideal method. An important limitation of the currently available methods including Lee's method is that, most of their predictive accuracy has not been validated against experimentally measured 'intracellular' fluxes. Considering that the major purpose of developing these methods is to accurately predict system-level and context-specific intracellular metabolic flux distribution, it would be better if existing or new methods prove how accurately they predict intracellular metabolic distribution by comparing their results with *in vivo* intracellular flux data.

Importantly, the most practical method does not guarantee the best or the most accurate method. The choice of the most appropriate method would depend on various factors such as biological systems of interest, primary objective of study, and the availability of experimental data. For instance, if we study fast-growing microorganisms such as *E. coli* and *S. cerevisiae* of which the assumption of biomass flux maximization in FBA has successfully predicted metabolic behavior, using the methods such as E-Flux and Åkesson's method that need *a priori* knowledge of an appropriate objective function of the system would not be a problem. However, in order to study a broad range of systems including microorganisms with variable biomass composition, pathogens in dormancy or in latent phase, or cells of a multi-cellular organism, the methods such as Lee's method and iMAT whose objective functions can be universally applied to a variety of conditions are more desirable for such practical applications. In addition, if we focus on examining clear changes in metabolic behavior of a system, and want to avoid data normalization issues, using the methods that require binarized gene expression data would be appropriate. However, if we need to see more finely grained information, and if it is hard to define the specific threshold above which the level of gene expression indicates physiological activeness of corresponding reactions, using the methods which incorporate continuous gene expression values would be useful. Lastly, although PROM is sorted as an impractical method in Table 1 mainly due to its requirements for a large number of experimental datasets with regulatory information, PROM identified knock-out phenotypes for *E. coli* and *M. tuberculosis* with accuracies as high as 95% [37]. Still, as a recent research paper shows that the methods that use multiple gene expression datasets does not necessarily give more accurate flux predictions [32], the requirement for a large amount of input data to perform the analysis, which might make the job more onerous, could be considered as another limitation of some of the existing methods from a practical point of view.

1.5 Objective of this study

The three main objectives of this study are:

(1) To develop new computational methods, for inferring system-level and condition-specific metabolic flux distributions from transcriptomic data, that overcome shortcomings of existing methods;

(2) To apply those methods to solve biological research problems, demonstrating their usefulness in biology;

(3) To test the generality of the usefulness of the methods through extensive validation with massive experimental data, and making the dataset used for validation publicly available for the research community in this field.

CHAPTER 2: Development of E-Flux2 and SPOT

2.1 Background

Integration of transcriptomics data in genome-scale metabolic models potentially enables the determination of context-specific system-wide metabolic fluxes through a relatively simple and low-cost omics technology.

There have been previous studies to integrate transcriptomic data with genome-scale metabolic models, which are covered in Chapter 1 of this dissertation and other reviews [12,14,18,30–32,57,58]. While powerful in many settings, existing methods have several shortcomings:

(1) Some of them require multiple sets of input data for a single analysis [37,38], which is often undesirable.

(2) Some methods require a user-defined threshold to define “high” or “low” expression states [42–44,59], which leads to subjective results since the specific threshold above which the level of gene expression indicates physiological activeness of corresponding reactions must be arbitrarily chosen and may vary across genes, conditions, or organisms.

(3) Several methods require *a priori* assumption of an appropriate objective function such as biomass production rate (i.e. the growth rate) [37,42,46,47]. The biomass flux is the most widely used objective function for FBA optimization problems [51]. Although the assumption of biomass flux maximization in FBA has successfully predicted metabolic behavior, especially of fast-growing microorganisms [53], we need a method which can be universally applied to a variety of organisms in cases where knowledge of the biological objective function is uncertain, such as microorganisms with variable

biomass composition, pathogens in dormancy or in latent phase, or cells of a multi-cellular organism [54].

(4) Several methods produce non-unique solutions. That is, they produce a solution out of a space of possibilities, all of which are in theory possible. If all solutions in the space of possibilities were equally good in terms of their ability to predict fluxes, this would be acceptable, but in general, there is a range of possibilities for predictive accuracy. A single solution that is arbitrarily chosen is difficult to reproduce and is typically dependent on the software or hardware used for the analysis [60]. Thus, if a method has non-unique solutions, a deterministic method to pick one of the good solutions (i.e. one of the ones with high predictive accuracy) is desirable.

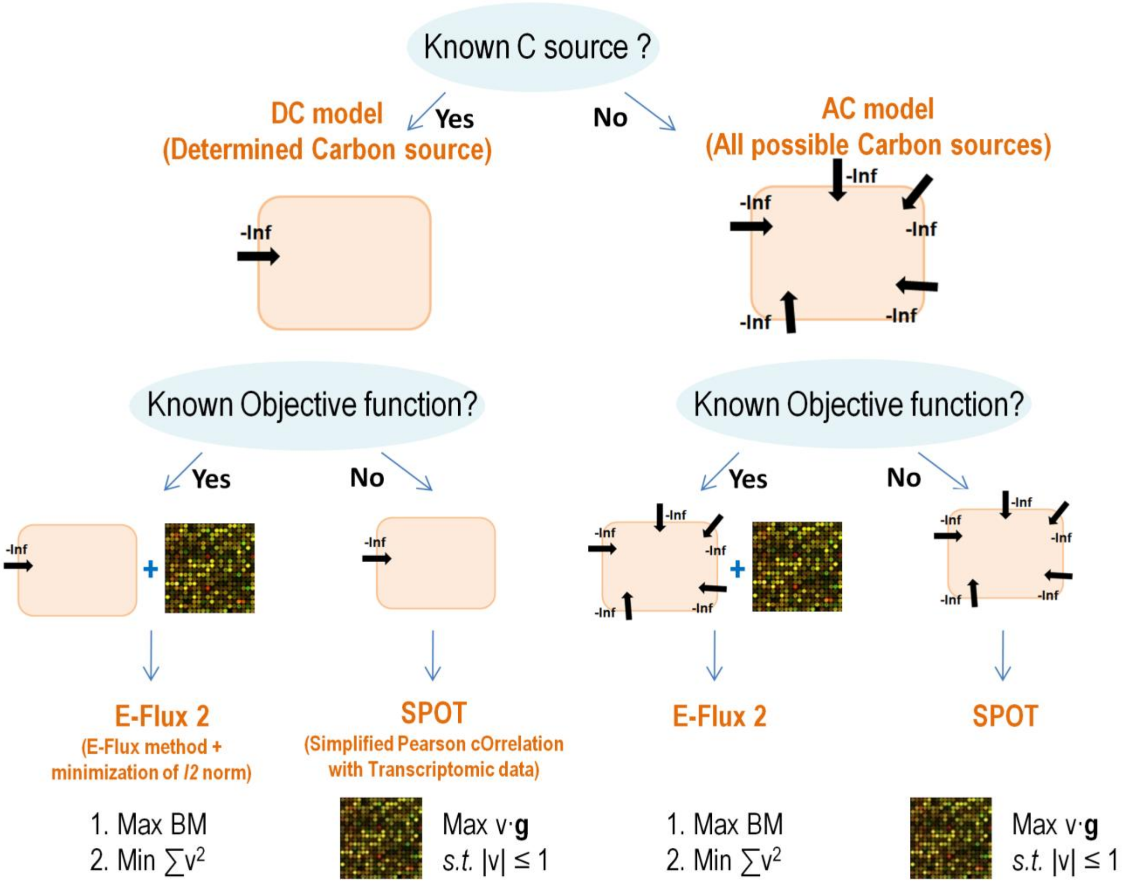
(5) Lastly, previous studies have generally focused on conditions where the carbon source of the system and its uptake rate are known. While many biotechnological and laboratory processes operate on a known single carbon source, typically glucose, we would sometimes like to study microorganisms living in conditions where the carbon source is unknown, such as for *in vivo* applications [61].

Therefore, in light of various experimental and cellular conditions in practical applications, there is still a need for a method that can provide all five of the desirable features listed in Table 2.1. Moreover, until recently, the predictive accuracy of previous methods had not been tested against experimentally measured intracellular fluxes [32]. It is, thus, unclear which method has the best accuracy in general because of limited validation.

	Desirable features	Benefits
1	Requirement for only a single gene expression data as input	Simpler analysis with less effort and cost
2	Use of continuous gene expression values without using arbitrary thresholds	Acquisition of more fine-grained information by avoiding arbitrary classification of gene expression levels
3	Capability to be used when an appropriate objective function is unknown	Applications to microorganisms with variable biomass composition, pathogens in dormancy or in latent phase, or cells of a multi-cellular organism
4	Capability to produce a unique metabolic flux distribution	More reproducible analysis independent of hardwares and softwares used to solve optimization problems
5	Capability to be used when the carbon source of the system and its uptake rate is unknown	Applications to microorganisms living in intact tissues or in natural environments

Table 2.1 Summary of the desirable features of a method for predicting intracellular metabolic fluxes using transcriptomic data-integrated genomic models. Five desirable features of a new method are listed in the left side of the table, and the corresponding benefits are described in the right column.

In this study, we compiled the most extensive dataset to date, consisting of 20 experimental conditions (11 in *E. coli* and 9 in *S. cerevisiae*, see Table 2.2 for details), of genome-wide gene expression measurements coupled with corresponding central carbon metabolism intracellular flux measurements. We used this dataset to rigorously evaluate the performance of representative methods for predicting intracellular metabolic fluxes using transcriptomic data. Based on this evaluation, we propose two new methods, E-Flux2 and SPOT, to be employed when a suitable biological objective is available and unavailable, respectively (Fig 2.1). The combination of the two methods provides a general strategy for predicting intracellular fluxes using transcriptomic data that satisfies all of the desirable features mentioned above. Depending on knowledge of the carbon source and availability of a suitable biological objective, this strategy achieves an average uncentered Pearson correlation of predictions against measurements over our dataset that ranges from 0.59 to 0.87, outperforming a representative selection of currently available methods.


Known C source ?
DC model
(Determined Carbon source)
Yes
No
AC model
(All possible Carbon sources)
-Inf
Known Objective function?
Known Objective function?
Yes
No
Yes
No
+
+
E-Flux 2
(E-Flux method + minimization of l2 norm)
SPOT
(Simplified Pearson cOrrelation with Transcriptomic data)
E-Flux 2
SPOT
1. Max BM
2. Min ∑v²
Max v·g
s.t. |v| ≤ 1
1. Max BM
2. Min ∑v²
Max v·g
s.t. |v| ≤ 1

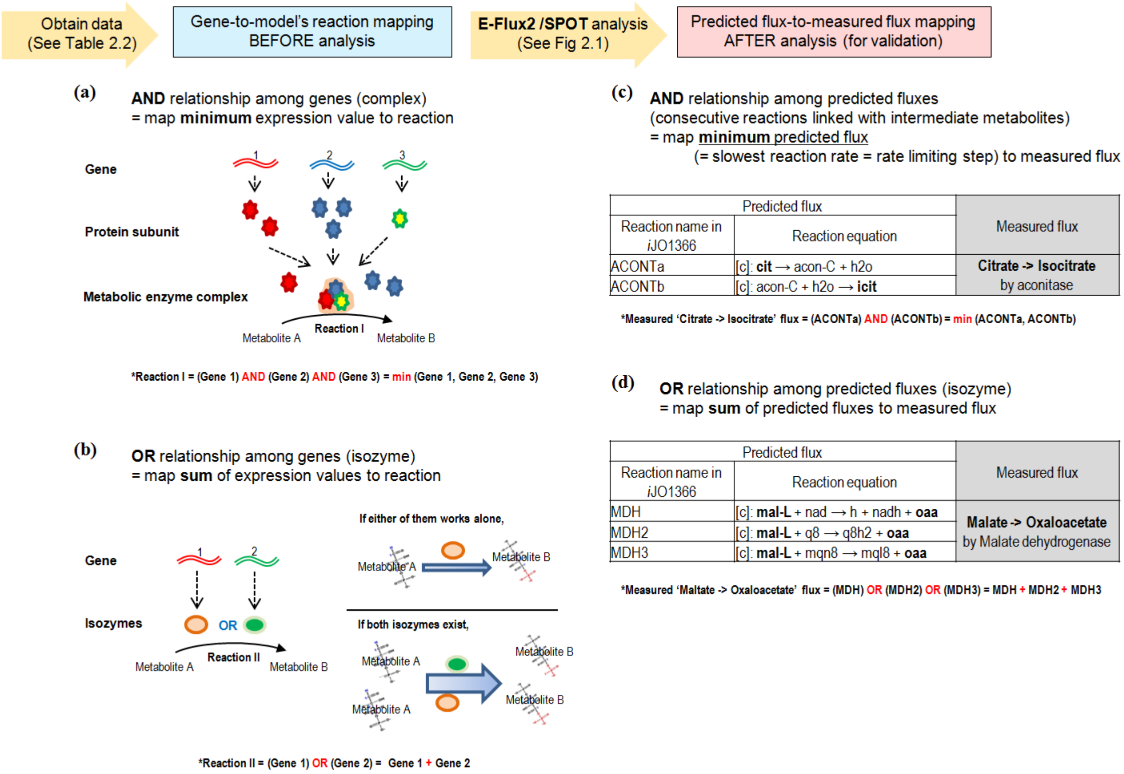
	Known C source	Unknown C source
Known Obj. func.	DC + E-Flux2	AC + E-Flux2
Unknown Obj. func.	DC + SPOT	AC + SPOT

Fig 2. 1. Flow chart illustrating how to choose between E-Flux2 and SPOT. If we know the cell's carbon source, we use the DC (determined carbon source) template model which has a negative infinity value on the lower bound of the known carbon source uptake reaction. Otherwise, we use the AC (all possible carbon sources) model which allows all carbon sources in the model to be taken up by the cell. If the biomass composition of the cell is known and the maximization of biomass flux is a suitable objective function, E-Flux2 (E-Flux method and

minimization of l2norm) can be used. Otherwise, we can use SPOT (Simplified Pearson cOrrelation with Transcriptomic data).

2.2 Materials and Methods

A description of the whole process for our research follows below. A schematic overview of it can be found in Fig 2.2.



(a) **AND** relationship among genes (complex)
= map **minimum** expression value to reaction

*Reaction I = (Gene 1) AND (Gene 2) AND (Gene 3) = min (Gene 1, Gene 2, Gene 3)

(b) **OR** relationship among genes (isozyme)
= map **sum** of expression values to reaction

*Reaction II = (Gene 1) OR (Gene 2) = Gene 1 + Gene 2

(c) **AND** relationship among predicted fluxes
(consecutive reactions linked with intermediate metabolites)
= map **minimum** predicted flux
(= slowest reaction rate = rate limiting step) to measured flux

Predicted flux		Measured flux
Reaction name in *i*JO1366	Reaction equation	
ACONTa	[c]: **cit** ⟶ acon-C + h2o	**Citrate -> Isocitrate** by aconitase
ACONTb	[c]: acon-C + h2o ⟶ **icit**	

*Measured 'Citrate -> Isocitrate' flux = (ACONTa) AND (ACONTb) = min (ACONTa, ACONTb)

(d) **OR** relationship among predicted fluxes (isozyme)
= map **sum** of predicted fluxes to measured flux

Predicted flux		Measured flux
Reaction name in *i*JO1366	Reaction equation	
MDH	[c]: **mal-L** + nad ⟶ h + nadh + **oaa**	**Malate -> Oxaloacetate** by Malate dehydrogenase
MDH2	[c]: **mal-L** + q8 ⟶ q8h2 + **oaa**	
MDH3	[c]: **mal-L** + mqn8 ⟶ mql8 + **oaa**	

*Measured 'Maltate -> Oxaloacetate' flux = (MDH) OR (MDH2) OR (MDH3) = MDH + MDH2 + MDH3

Fig 2.2. The process for our research classified into five steps. The steps are: 1) obtaining both transcriptomic and fluxomic data measured under the same conditions (See Table 2.2); 2) mapping gene expression data onto corresponding reactions in the model based on Gene-Protein-Reaction associations; 3) creating one of the two template metabolic models depending on carbon

source information; 4) solving an optimization problem with one of the two algorithms depending on the availability of information on biomass objective (See Fig 2.1); and 5) calculating the correlation between predicted and measured fluxes. Figs 2.2a and 2.2b illustrate how transcriptomic data were mapped onto corresponding reactions in the model based on Gene-Protein-Reaction associations. (a) In the case where multiple enzymes form a complex to mediate a certain metabolic reaction (AND relationship), we mapped the minimum value of the expression level of the associated genes encoding its subunits to the corresponding reaction since the least-expressed components is likely to determine the final concentration of the complete enzyme complex. (b) If a reaction is catalyzed by isozymes (OR relationship), we took the sum of the expression values of the associated genes for mapping since the total capacity of the reaction is given by the sum of the capacities of its isozymes. Figs 2.2c and 2.2d show how predicted fluxes were matched with a corresponding measured flux to calculate correlation between them. (c) If a certain measured reaction corresponds to a set of consecutive reactions in the model that share intermediate metabolites (AND relationship), the slowest reaction rate also known as the rate-determining step (the minimum flux) among those predicted fluxes was used to match with the corresponding measured flux. (d) In the case where a measured flux corresponds to multiple reactions in the model that mediate an identical chemical conversion independently with each other (OR relationship), the sum of those predicted fluxes was used to match with the corresponding measured flux.

2.2.1 Transcriptomic data, fluxomic data, and metabolic models used for this study

The first step was to collect a dataset of transcriptomic and fluxomic measurements obtained from cells under the same conditions. The measured fluxes were obtained to compare them with the predicted fluxes. To this end, we obtained data published by Ishii *et al.*[62] and Holm *et al.* [63] for *E. coli*, and by Rintalta *et al.* [64,65] and Celton *et al.* [9] for *S. cerevisiae*, where both expression data and ^{13}C flux data measured under the identical conditions can be acquired. The dataset is made up of total 20 experimental conditions (11 in *E. coli* and 9 in *S. cerevisiae*), a detailed description of which is given in Table 2.2.

		E. coli	*S. cerevisiae*
Genome-scale metabolic model	For Table 2.3 & most Figs	*i*JO1366 [66]	Yeast 5 [67]
	For Figs 2.5 and 2.6	*i*JO1366 [66], *i*AF1260 [68], *i*JR904 [69]	Yeast 5 [67], *i*MM904 [70], *i*ND750 [71]
Transcriptomic data & measured flux data	Dataset 1	Ishii *et al.*, 2007 [62] • Data were measured under 8 different conditions - wild type *E. coli* cells cultured at a growth rate of 0.2, 0.5, and 0.7 hours^{-1}, and single-gene knockout mutants (Δ*pgm*, Δ*pgi*, Δ*zwf*, Δ*rpe*, Δ*gapC*) • The transcriptomic data are two-color microarray data normalized using MAANOVA [72]. • The number of measured fluxes used for validation: around 248 measured fluxes (31 fluxes per condition, and total 8 conditions)	Rintala *et al.*, 2009 [64] - transcriptiomic data Jouhten *et al.*, 2008 [65] - fluxomic data • Data were obtained from yeast cells cultured in 5 different oxygen levels (20.9, 2.8, 1.0, 0.5, and 0.0% O_2) • The transcriptomic data are single-color microarray data normalized with Robust Multichip Average normalization [73] • The number of measured fluxes used for validation: 110 fluxes (22 fluxes per condition, and total 5 conditions)
	Dataset 2	Holm *et al.*, 2010 [63] • Data were obtained from 3 different *E. coli* strains that	Celton *et al.*, 2012 [9] • Data were collected from yeast cells treated with 4 different

		are wild-type cells, NADH oxidase- overexpressing cells, and the soluble F1-ATPase-overexpressing cells • The transcriptomic data are single-color microarray data normalized using the Qspline method [74] • The number of measured fluxes used for validation: 66 fluxes (22 fluxes per condition, and total 3 conditions)	concentrations of acetoin (0, 100, 200, and 300 mM) • Its transcriptomic data are two-color microarray data normalized using MAANOVA [72] • The number of measured fluxes used for validation: 116 fluxes (29 fluxes per condition, and total 4 conditions)
	Total	11 conditions in *E. coli*	9 conditions in *S. cerevisiae*

Table 2.2. Datasets and metabolic models used for this study. We used the experimental datasets published in the papers listed in the lower row, each of which has both transcriptomic data and fluxomic data measured under same condition. We needed experimentally measured fluxes data to validate predictive accuracy of our methods by comparing them with the predicted fluxes.

As the metabolic models for *E. coli* and *S. cerevisiae*, we used *i*JO1366 [66] and Yeast 5 [67], respectively, in most tables and figures of this paper. As shown in Fig 2.5, we also tested our methods on older models of *E. coli* (*i*JR904 [69] and *i*AF1260 [68]) and of *S. cerevisiae* (*i*ND750 [71] and *i*MM904 [70]) to examine the applicability of our methods to the relatively incomplete models. The model files, and the transcriptomic and fluxomic datasets that were used in this study are given in S1 Dataset, and S2 Dataset, respectively at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157101#sec021.

2.2.2 Creation of template metabolic models depending on carbon source information

When integrating transcriptomic data with genome-scale metabolic models, a problem of scaling can occur because the units for measuring metabolic flux and the units for measuring gene expression are not related. For instance, if the carbon uptake rate is set to 1, and the transcriptome values are all in the order of 10000, then applying such values as upper bounds will not constrain the model. To avoid this issue, we construct a template model that is independent of *a priori* information on cellular uptake rates and ATP maintenance flux. The template model is made by setting the flux bounds either to zero or to positive or negative infinity while maintaining the stoichiometric and reversibility information of the original genome-scale model:

$$a_j \le v_j \le b_j \;\; \rightarrow \;\; \overline{a_j} \le v_j \le \overline{b_j} \tag{2}$$

where, for all j,

$$\overline{a_j} = \begin{cases} 0, \text{if } a_j \ge 0, \\ -\infty, \text{if } a_j < 0, \end{cases} \text{ and } \overline{b_j} = \begin{cases} +\infty, \text{if } b_j > 0, \\ 0, \text{if } b_j \le 0, \end{cases}$$

where v is a flux vector representing the reaction rates of the n reactions in the network, and a_j and b_j are the minimum and maximum reaction rates through reaction j defined in the original model. In this manner, we constructed two kinds of template models to simulate two different situations depending on whether we know which carbon source the cell uses. One template model, which we call 'DC (determined carbon source)', has a lower bound of negative infinity for the known carbon source uptake reaction. The other one, which we call 'AC (all possible carbon sources)', allows all carbon sources in the model to be taken up by the cell. Among all metabolites participating in the exchange reactions, the set of possible carbon sources were selected based on their chemical formula. The list of carbon sources whose uptake rate were set as negative infinity in the AC models for both microorganisms are given in S1 Table of Supporting Information. Inorganic metabolites such as ions and water molecules were allowed to be taken up by the DC and AC models if their original genome-scale metabolic models did so. The information pertaining to each specific model we used can be found in S1 Dataset of Supporting information.

This step, converting original genome-scale models into DC or AC template models before integrating gene expression data, resolves the scaling problem described above. The fluxes predicted by our method have an arbitrary unit. Thus, the relative magnitude of predicted fluxes across reactions is meaningful, but their absolute magnitude is not. Any known or measured reaction rate (e.g. glucose uptake rate, ATP maintenance flux, and oxygen uptake rate that are discarded when building a DC or AC template model) can be used to normalize the predicted fluxes to an absolute reference.

2.2.3 Two different optimization strategies depending on the availability of biomass objective

If information on the biomass composition of a certain organism is available and maximizing its growth rate is appropriate for prediction, our first method, called E-Flux2, is an effective way to study its metabolic behavior. Otherwise, our second method, called SPOT, can be used.

2.2.3.1 E-Flux2

E-Flux is an extension of FBA that infers a metabolic flux distribution from transcriptomic data [46,47]. The rationale behind E-Flux is that, given a limited translational efficiency and a limited accumulation of enzyme over the time, the mRNA level can be used as an approximate upper bound on the maximum amount of metabolic enzymes, and hence as a bound on reaction rates. The standard FBA involves solving the following linear optimization problem:

$$\max\ f'v$$
$$\text{subject to} \begin{cases} Sv = 0 \\ \overline{a}_j \le v_j \le \overline{b}_j \end{cases} \qquad (2)$$

where f is a coefficient vector defining the organism's objective function, S is the stoichiometric matrix.

The main difference between E-Flux (equation (3) below) and the standard FBA (equation (2) above) is that E-Flux uses g_j, the absolute gene expression level associated with reaction j, for an upper bound, b_j^e, and sets a lower bound, $a_j^e = -g_j = -b_j^e$ for reversible reactions, otherwise $a_j^e = 0$. Here, absolute gene expression refers to any transcript abundance measurement in arbitrary units.

For one-color microarrays and RNA-seq measurements, it is relatively straightforward to determine absolute gene expression [75]. For two-color microarrays, however, it is more difficult to determine absolute gene expression because of effects such as spot size variation, and relative expression between two conditions is typically reported [76]. It is, however, possible to normalize two-color microarray data so that the gene expression levels can be compared both within an array and across arrays by estimating and removing non-biological effects, such as dye-specific, spot-specific, and array-specific effects [77,78]. For two-color microarray data (i.e. the datasets from Ishii *et al.* and Celton *et al.*), we used the MAANOVA normalization method [72] to achieve this normalization. MAANOVA uses an ANNOVA model to estimate and remove non-biological effects. We have previously used this method to successfully obtain estimates of absolute gene expression from two-color microarray measurements for E-Flux [46].

E-Flux solves the following:

$$\max\ f'v$$
$$\text{subject to}\begin{cases} Sv = 0 \\ a_j^e \le v_j \le b_j^e \end{cases} \tag{3}$$

where, for all j,

$$a_j^e = \begin{cases} -g_j, \text{if } \overline{a_j} < 0, \\ 0, \text{if } \overline{a_j} \ge 0, \end{cases} \text{ and } b_j^e = \begin{cases} g_j, \text{if } \overline{b_j} > 0, \\ 0, \text{if } \overline{b_j} \le 0. \end{cases}$$

Gene expression data were mapped to corresponding reactions in the network based on gene-protein-reaction (GPR) associations. For example, in the case where an enzyme complex consisting of subunits encoded by multiple genes mediates a certain metabolic reaction, we put the minimum value of the expression level of the associated genes on b_j^e because the minimum concentration of the components determines the maximum concentration of the complete enzyme complex (see Fig 2.2a). If a reaction is catalyzed by isozymes, we took the sum of the expression values of the associated genes for b_j^e since the total capacity of the reaction is given by the sum of the capacities of its isozymes (Fig 2.2b). If either the gene expression or GPR association relationship is unavailable for a certain reaction, then the values of a_j^e and b_j^e of that reaction were kept as they were defined in the template model (0 or positive or negative infinity, see equation (1) above) so as not to constrain the model unnecessarily.

A problem of E-Flux is that the outcome solution is not unique, making it difficult to clearly identify predicted metabolic responses. Bonarius *et al.* [79] used minimization of the Euclidean norm as an objective function to find a unique metabolic flux distribution in hybridoma cells. The Euclidean norm of a vector x, also called the l^2 norm, is given by:

$$|x|_2 = |x| = \sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_n}^2} \text{ for } x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{4}$$

whose intuitive geometric meaning is the length of a vector x on an n-dimensional Euclidean space R^n. Thus, as stated by Bonarius *et al.*, the constraint of minimizing the Euclidean norm corresponds to the strategy of a cell to minimize the length of the metabolic flux vector to channel metabolites as efficiently as possible. We additionally applied this theoretical constraint after maximizing the biomass flux to find a unique metabolic flux distribution satisfying both optimal biomass flux and the flux minimizing its Euclidean norm. So, the first method that we propose, which we call "E-Flux2" (meaning E-Flux method combined with minimization of l^2 norm), consists of two steps of optimization, which can be chosen when a suitable objective function is known:

Step 1. E-Flux

Step 2. Minimization of l^2 norm

$$z^* = \max f'v \quad \text{subject to} \begin{cases} Sv = 0 \\ a_j^e \le v_j \le b_j^e \end{cases} \quad \rightarrow \quad \min \sum_{j=1}^{n} {v_j}^2 \quad \text{subject to} \begin{cases} Sv = 0 \\ a_j^e \le v_j \le b_j^e \\ f'v = z^* \end{cases} \tag{5}$$

After calculating the optimal biomass flux, denoted as z^* here, this method finds a unique metabolic flux distribution by minimizing the Euclidean norm of the flux vector. The square root function was ignored since removing the square root does not change the solution. Since the objective function, the Euclidean norm squared, is strictly convex, and all equality and inequality constraints are linear, which is convex, the solution of E-Flux2 is unique because the optimal solution to the problem of minimizing a strictly convex function over a convex set is unique [80]. The output vector calculated by E-Flux2 can be biologically interpreted as a metabolic flux distribution that allows the cell to achieve maximum growth rate in an energy efficient way. The idea underlying E-Flux2 is similar to parsimonious FBA (pFBA) in which FBA is followed by minimization of the l^1 norm (or Manhattan norm) [81]. pFBA does not, however, necessarily produce a unique solution since the objective function, the l^1 norm, is not strictly convex.

Though not largely different from E-Flux, E-Flux2 overcomes a major shortcoming of E-Flux, namely, that it does not yield a unique solution. Among the space of solutions that E-Flux provides, E-Flux2 provides a method to select one solution in a manner that is intuitive and yields high correlation to measured fluxes (see sections 2.3 and 2.4).

2.2.3.2 SPOT

If a suitable objective such as the biomass flux is unknown, we can use a second method which is to maximize correlation between a flux vector, v, and its corresponding gene expression data, g. The assumption behind this strategy is that enzymatic transcript concentrations and metabolic fluxes can be related to each other, albeit in a complex manner, since the existence of a transcript is necessary for the presence or activity of its corresponding enzyme [49]. To calculate the correlation, we used the uncentered Pearson product-moment correlation which is a popular measure of the linear correlation between two variables, resulting in the following optimization problem:

$$\max \frac{v \cdot g}{\|v\| \|g\|} = \frac{\sum_{j=1}^{n} v_j g_j}{\|v\| \|g\|} \qquad (6)$$

$$\text{subject to} \begin{cases} Sv = 0 \\ \overline{a_J} \leq v_j \leq \overline{b_J} \end{cases}$$

where, for all $j = 1, \ldots, n$,

$$g_j = \begin{cases} g_j, \text{if } g_j \text{ is available and } \overline{a_J} \geq 0, \\ -g_j, \text{if } g_j \text{ is available and } \overline{b_J} \leq 0, \\ 0, \quad \text{otherwise} \end{cases}$$

we will consider a problem with modified upper and lower bounds that are 0 or ± infinity as described in equation (1). If the network contains reversible reactions, the objective function of problem (6) is potentially problematic because the directions of reversible reactions (signs of their fluxes) are unknown, while gene expression is always positive as shown in Fig 2.3a

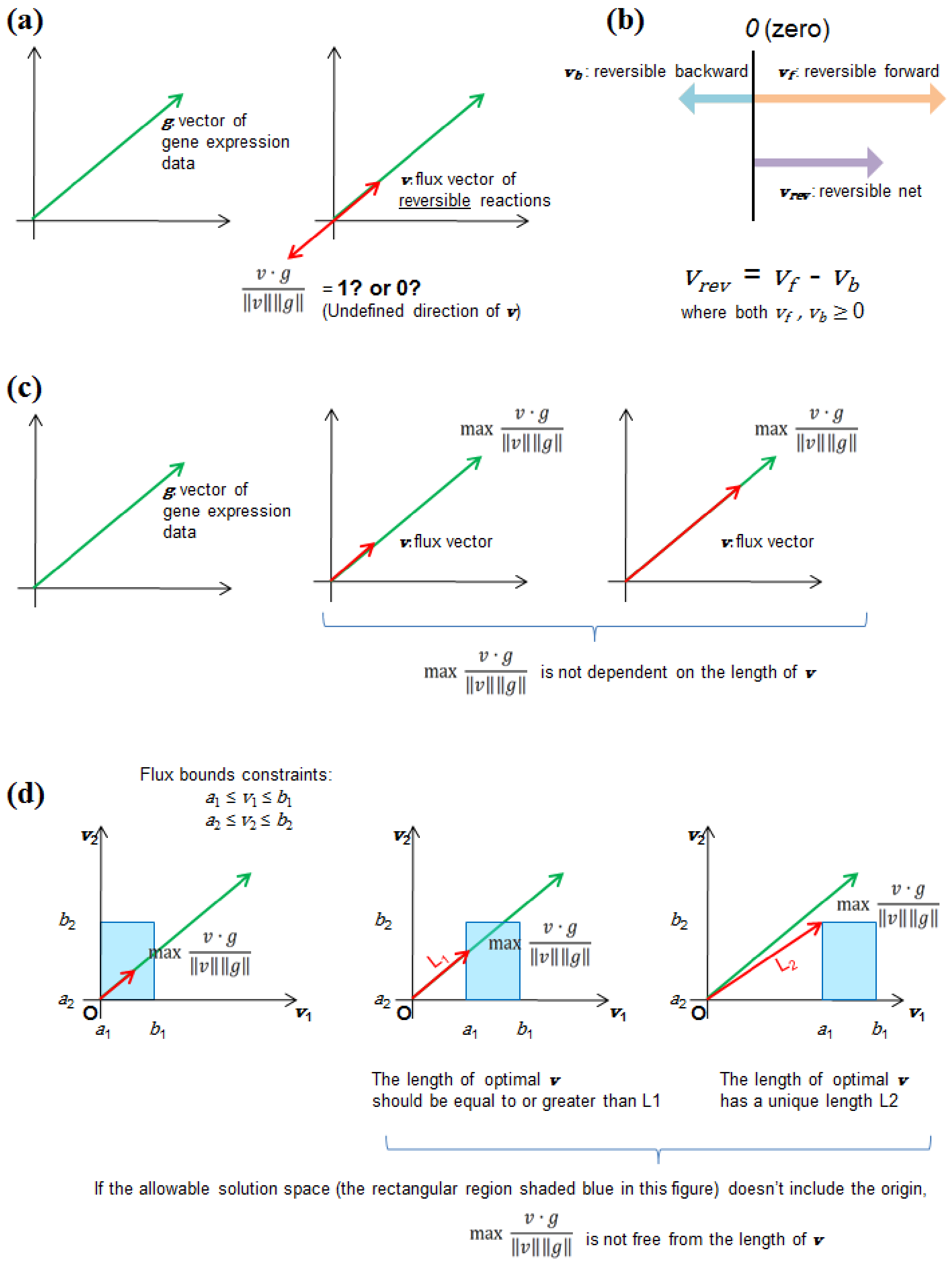

(a)
g: vector of gene expression data
v: flux vector of reversible reactions
$\frac{v \cdot g}{\|v\|\|g\|}$ = 1? or 0?
(Undefined direction of v)
(b)
0 (zero)
v_b: reversible backward
v_f: reversible forward
v_{rev}: reversible net
$V_{rev} = V_f - V_b$
where both v_f, $v_b \geq 0$
(c)
g: vector of gene expression data
v: flux vector
max $\frac{v \cdot g}{\|v\|\|g\|}$
max $\frac{v \cdot g}{\|v\|\|g\|}$ is not dependent on the length of v
(d)
Flux bounds constraints:
$a_1 \leq v_1 \leq b_1$
$a_2 \leq v_2 \leq b_2$
v_2
v_1
b_2
a_2
O
a_1
b_1
L1
L2
The length of optimal v should be equal to or greater than L1
The length of optimal v has a unique length L2
If the allowable solution space (the rectangular region shaded blue in this figure) doesn't include the origin,
max $\frac{v \cdot g}{\|v\|\|g\|}$ is not free from the length of v

Fig 2.3 Rationale for the SPOT method. (a) We noticed that the dot product between a flux vector (v, denoted as red arrows in the figure) and its corresponding gene expression data (g, denoted as green arrows in the figure) in the numerator of the objective function cannot be calculated for the set of reversible reactions since the directions of reversible reactions are

undefined whereas gene expression data values are always positive. (b) So we decomposed every reversible reaction in the model into two positive irreversible reactions, the forward reaction, v^f, and the backward reaction, v^b, where $v^{rev} = v^f - v^b$, and $v^f, v^b \geq 0$ (c) Usually, the maximum Pearson product-moment correlation is not dependent on the length of v but dependent on the angle between v and g. (d) However, if any of the flux bounds does not include zero, the origin in the graph, the maximum correlation is no longer independent of the length of the flux vector, v. Thus, it is a prerequisite for using SPOT method to make sure that the allowable solution space includes the origin. The light blue-colored rectangular space shows the allowable solution space that is determined by flux bounds (i.e. a_1, b_1 and a_2, b_2 in the figure) of all reactions (i.e. v_1 and v_2).

We therefore decomposed each reversible reaction j into two positive irreversible reactions, the forward reaction, v_j^f, and the backward reaction, v_j^b, where $v_j = v_j^f - v_j^b$, and $v_j^f, v_j^b \geq 0$ (Fig 2.3b). Let us assume without loss of generality that reactions $1, \dots, n-r$ are irreversible, while reactions $n-r+1, \dots, n$ are reversible, and that all irreversible reactions are defined in the forward reactions (i.e. their fluxes are non-negative). Then, instead of problem (6), we solve:

$$\max \frac{\bar{v} \cdot \bar{g}}{\|\bar{v}\|\|\bar{g}\|} = \frac{\sum_{k=1}^{n-r} v_k g_k + \sum_{k=n-r+1}^{n} v_k^f g_k + \sum_{k=n-r+1}^{n} v_k^b g_k}{\|\bar{v}\|\|\bar{g}\|} \tag{7}$$

$$\text{subject to} \begin{cases} \bar{S}\bar{v} = 0 \\ v \geq 0 \\ v^f \geq 0 \\ v^b \geq 0 \end{cases}$$

where $\bar{v} = \left[v^{irr}\ v^{f}\ v^{b}\right]^{T} = \left[v_1 \cdots\ v_{n-r}\quad v^{f}_{n-r+1} \cdots v^{f}_{n}\quad v^{b}_{n-r+1} \cdots v^{b}_{n}\right]^{T}$, $\bar{g} = \left[g^{irr}\ g^{rev}\ g^{rev}\right]^{T} = [g_1 \cdots\ g_{n-r}\quad g_{n-r+1} \cdots g_n\quad g_{n-r+1} \cdots g_n]^{T}$, $\bar{S} = [S^{irr}\quad S^{rev}\quad -S^{rev}\ ]$, S^{irr} is the submatrix consisting of the first $n-r$ columns of S, and S^{rev} is the submatrix consisting of columns $n-r+1$ to n of S. Solving this optimization problem is computationally inefficient since the form of the objective function is nonlinear. However, this problem can be converted to an equivalent semi-definite programming problem (8):

$$\max\ \bar{v} \cdot \bar{g} \tag{8}$$

$$\text{subject to} \begin{cases} \bar{S}\bar{v} = 0 \\ 0 \leq\ \bar{v} \\ ||\bar{v}||^2 \leq 1 \end{cases}$$

This is the second method we propose, which we call "SPOT" (Simplified Pearson cOrrelation with Transcriptomic data). SPOT can be used when biomass flux is not an appropriate optimization objective. The conversion of optimization problem (7) to SPOT (8) is based on a few steps of justification.

First, the maximum Pearson product-moment correlation is not dependent on the length of the flux vector, $\bar{v}$ (see Fig 2.3c and Appendix 1). Thus, the norm of $\bar{v}$ can be ignored in the objective function. Since the norm of $\bar{g}$ is a constant that only affects the objective value, not the optimal flux distribution, it was also removed in the objective function. Lastly, to avoid the situation where the maximum value of $\bar{v} \cdot \bar{g}$ goes to infinity, the norm of $\bar{v}$ was constrained to an arbitrary number, in this case, 1.

The optimization problem described in equation (7) can be simplified to SPOT only if the maximum correlation is independent of the length of the flux vector, $\bar{v}$. This is true provided the allowable flux solution space includes the origin, which is indeed the case.

Fig 2.3d explains this geometrically. The solution of (8) is unique (see Appendix 2 for the proof).

Transcriptomic data are used to constrain fluxes in the model for E-Flux2, and they are used to define the objective function for SPOT. The process of making a choice between E-Flux2 and SPOT is described in the flow chart in Fig 2.1.

2.2.4 Validation of the predictive accuracy of the algorithm using the measured fluxes

The predictive accuracy of our algorithm was validated by calculating the uncentered Pearson product-moment correlation between *in silico* fluxes and corresponding ^{13}C-determined *in vivo* intracellular fluxes, that is

$$\frac{v_p \cdot v_m}{\|v_p\| \|v_m\|} \tag{9}$$

where v_p and v_m are the predicted and measured vectors of intracellular fluxes, respectively, and $\|\cdot\|$ denotes the l^2 norm. The uncentered Pearson correlation is a good metric of the performance of flux inference methods because these methods allow determination of fluxes only within an unknown scale factor. A value of the correlation coefficient close to +1 or -1 indicates a strong positive or negative linear relationship between v_p and v_m, respectively. A value of 0 indicates no linear relationship [82].

We found that some of the measured fluxes are not directly matched with predicted fluxes of the model in a 1-to-1 relationship since the reactions described in the model are more detailed. Like the GPR association relationships that were used to match genes with corresponding reactions, we identified OR or AND relationships between predicted fluxes (Figs 2.2c and 2.2d). If a measured reaction corresponds to the set of consecutive reactions in the model that are linked with intermediate metabolites (AND relationship, Fig 2.2c), then the minimum flux value—the slowest reaction rate—among those predicted fluxes was used to calculate correlation with the corresponding measured flux since the rate of a reaction with several steps is determined by the slowest step, which is known as the rate-limiting step in chemical kinetics [83]. If a measured flux corresponds to multiple identical reactions (OR relationship, Fig 2.2d), the sum of those predicted fluxes was used to calculate the correlation since the rate of a reaction would be faster, that is, would have greater flux value, as the number of reactions that can perform an identical chemical conversion increases.

The reactions whose measured fluxes were used to calculate the correlation for each dataset are shown in S2 Dataset of Supporting Information. It should be noted that our validation is directly based only on these reactions and, in general, they belong to central carbon metabolic pathways. We hypothesize that our flux predictions for other reactions (e.g., reactions in secondary metabolism) are likely also to be good, given the interconnected nature of metabolism, but our data do not allow us to directly test this hypothesis. Another thing to note is that all data were gathered from cells grown on glucose. There are, therefore, significant similarities among all the measured flux distributions, and indeed, it is possible to find a single flux distribution for *E. coli* and a single flux distribution for *S. cerevisiae* that each achieve high correlations with the measured data in each organism (data not shown). Nevertheless, the dataset we have gathered is the largest and most comprehensive dataset that currently exists for validating methods of predicting intracellular fluxes from transcriptomic data. We expect that the high correlations obtained by E-Flux2 and SPOT will generalize beyond *E. coli* and *S.*

cerevisiae growing on glucose, given how their underlying optimizations reflect our general understanding of the relationship between metabolic flux and gene expression, but we cannot conclude this without additional data. In particular, coupled trascriptomic and fluxomic data obtained in organisms under very different conditions (e.g., organisms growing photoautotrophically or organisms under non-growth conditions) would help significantly in establishing the generality of our method.

2.2.5 Algorithm implementation of our methods

All methods in this study initially implemented in MATLAB (The Mathworks, Inc., Natick, Mass.). These were tested using MATLAB R2013b with Gurobi Optimizer 5.6 (Gurobi Optimization, Inc., Houston, Texas). SBMLToolbox was used to convert an SBML (Systems Biology Markup Language) model into a MATLAB data structure [84]. Computations were carried out on the Window 8 platform using a personal computer with an Intel Core i5 3.10 GHz processor with 8 GB of RAM. E-Flux2 and SPOT methods are also implemented in a freely downloadable software package called MOST (Metabolic Optimization and Simulation Tool) which is available at http://most.ccib.rutgers.edu/ whose source code is open to the public [85].

2.3 Results

2.3.1 Validation of the accuracy of our predictions against measured intracellular fluxes

The Pearson correlation between the predicted and the measured intracellular fluxes was calculated to validate the predictive accuracy of our method. All correlation values used to draw figures and tables are summarized in S2 Table of Supporting Information. The correlation values were grouped into four different cases depending on the availability of carbon source or objective function information. Biomass flux and glucose were used as the known objective function and the known carbon source in this study. The bold number in each category of the table presents the average correlation of 11 samples in *E. coli* and 9 samples in *S. cerevisiae*. The number on the right side of the plus minus sign indicates its standard deviation.

As summarized in Table 2.3, overall, the predicted fluxes of our method showed good correlation with the measured fluxes both in *E. coli* and *S. cerevisiae*. The result implies that our method can predict the measured intracellular fluxes best when we have knowledge of both carbon source and objective function (DC+E-Flux2, average correlation: 0.8683). Our algorithm is able to predict intracellular metabolic fluxes with a good correlation if the information on either biomass objective or carbon source is unknown as we can see in the category of DC+SPOT (average correlation: 0.8030), and AC+E-Flux2 (average correlation: 0.6733). In the case where there is no information on both carbon source and biomass objective, our AC+SPOT method allows us to predict intracellular metabolic fluxes with an average correlation of 0.5927. Although this value is weaker than those of the other three cases of our method, a Pearson correlation coefficient around 0.6 nevertheless represents moderate positive correlation [86].

	Known C source (glucose, in this case)					Unknown C source	
Known objective function (biomass, in this case)	Standard FBA$^{1)}$	pFBA$^{1)}$	FBA+min l^2	DC+E-Flux$^{1)}$	**Our method** (DC+E-Flux2)	AC+E-Flux$^{1)}$	**Our method** (AC+E-Flux2)
	Known C uptake rate			Unknown C uptake rate			
	[0.4965, 0.8516] **0.7952** ± 0.2317	**0.8337** ± 0.1800	**0.8106** ± 0.1740	[0.3506, 0.9223] **0.7829** ± 0.1307	**0.8683** ± 0.0964	[0.0027, 0.8625] **0.4516** ± 0.2343	**0.6733** ± 0.1349
Unknown objective function	DC+Lee *et al.*			**Our method** (DC+SPOT)		AC+Lee *et al.*	**Our method** (AC+SPOT)
	0.5792 ± 0.3642			**0.8030** ± 0.0342		**0.1257** ± 0.1268	**0.5927** ± 0.0974

Table 2.3 Validation of the accuracy of our predictions against measured intracellular fluxes. The Pearson correlation between the predicted and the measured intracellular fluxes was calculated to validate the predictive accuracy of our method. The correlation values were grouped into four different cases depending on the availability of carbon source or objective function information. The bold number in each category of the table presents the average correlation of 11 samples in *E. coli* and 9 samples in *S. cerevisiae.* The number to the right of the ± indicates its standard deviation. Since the fluxes predicted by standard FBA, pFBA and E-Flux are not unique, the output flux obtained using our specific implementation was used to calculate the average correlation. For FBA and E-Flux solutions, the minimum and the maximum correlations between predicted fluxes and the measured fluxes that each method can theoretically achieve are given

within square brackets above their average correlations. The way that we calculated the possible range of correlations of each method is described in Supplementary Methods in S1 Supporting Information. Note that the maximum possible correlation can be calculated only when we already have the measured flux datasets. There is no way to force each method to produce a metabolic flux distribution that achieves the best correlation with the measured fluxes. Our methods, E-Flux2 and SPOT, were developed during the process of testing various strategies for producing unique flux distributions and identifying those that achieve good correlation on average with measured fluxes.

(1) metabolic flux distributions produced by these methods - FBA, pFBA, E-Flux- are not unique

To see whether the good or modest correlation value between predicted and measured fluxes in each case is because of a good correlation between transcripts and measured fluxes itself, we also calculated the correlation between gene expression data and measured fluxes. When calculating the correlation with gene expression data, we used the absolute values of measured metabolic fluxes since gene expression values are always positive, while we used the signed measured fluxes when calculating the correlations with predicted fluxes. The correlation between gene expression data and the absolute values of the measured fluxes was 0.4923 (standard deviation: 0.2900), which is weaker than all of the correlations between the predicted fluxes and the measured fluxes. Although this value cannot be directly compared with correlations within the table (since they are calculated differently), this relatively poor correlation between gene expression data and the measured fluxes, given as a point of comparison, suggests that the correlation is improved by incorporating the gene expression data into a genome scale model.

2.3.2 Comparison of correlation with competing methods

Using the same transcriptomic and fluxomic datasets, we compared the accuracy of our predictions with other competing methods. We chose to compare against E-Flux [46,47] and the approach of Lee *et al.* [48], which are representative of competing methods which use a single transcriptomic dataset for an analysis without thresholds. Moreover, these two methods were compared against other methods of a similar nature including GIMME [42] and iMAT [43,44], and showed better performance in predicting exometabolome fluxes [48] or in robustness analysis [32]. For the Lee *et al.* method, we used an implementation provided with the original publication.

In all four scenarios with varying availability of carbon source or biomass objective information, our method outperformed existing methods in that it showed a higher average correlation with a smaller standard deviation (Table 2.3). Particularly when the carbon source is known but the biological objective is unknown, the Lee et al. method gives better predictions in *E. coli* (average correlation: 0.8887) but worse predictions in yeast (average correlation: 0.2009) than our method (DC+SPOT) whose average correlation is 0.7960 and 0.8117 in *E. coli* and in yeast, respectively (S2 Table of Supporting Information). Unlike a prokaryote model such as *i*JO1366, a eukaryote model such as Yeast 5 is compartmentalized into organelles (e.g. mitochondria, peroxisomes, lysosomes, ER, and nucleus). As can be seen in S2 Dataset of Supporting Information, the set of measured intracellular fluxes that were used for validation includes interorganelle transport reactions such as pyruvate transport between cytoplasm and mitochondria where the incorrect predictions of the Lee et al. method mainly occurred. Considering the importance of compartmentalization in eukaryotic metabolic models [87], our method seems to be more desirable to study more complex systems since it is less influenced by whether the model is compartmentalized or not.

We also have carried out standard FBA and parsimonious FBA (pFBA) for reference [81]. pFBA was performed using the COBRA Toolbox [88]. Since standard FBA and pFBA require *a priori* information on several specific fluxes such as sugar (e.g. glucose) uptake rate and oxygen exchange rate, these fluxes were set according to the experimental conditions described in the four papers where the transcriptomic and fluxomic data sets were obtained. Simulating anaerobic growth with Yeast 5 requires the simulated medium to be supplemented with phosphadiate and sterols and modification of the biomass definition [67]. Due to inconsistency with the experimental condition, we could not evaluate the performance of standard FBA and pFBA in the 0% oxygen condition of the Rintalta *et al.* dataset (S2 Table of Supporting Information).

Since standard FBA and pFBA need both carbon source and objective function information, their correlations can be compared with those of E-Flux and E-Flux2 in Table 2.3. Our method (0.8683, SD: 0.0964) performs better than standard FBA (0.7952, SD: 0.2317) and pFBA (0.8337, SD: 0.1800) in terms of both the correlation and the standard deviation. In the previous study by Machado and Herrgård [32], pFBA has been shown to have a higher overall predictive capability over various methods that integrate gene expression data, which casts doubt on the necessity of utilizing transcriptomic data in constraint-based modeling. Our result, however, suggests that integration of gene expression data can be used to improve flux distribution predictions, particularly when the carbon and oxygen uptake rates are unknown.

Importantly, the result predicted by our method is unique. The Lee et al. method also produces a unique solution using geometric FBA [60], which identifies the center of a solution space. Since halfway between infinity and zero or between plus and negative infinity is not defined, we set lower and upper bounds of the models to either zero or ±1000 (1000 is chosen as an arbitrary, large number) to run the Lee et al. method. For standard FBA and E-flux, which do not necessarily produce a unique flux distribution, and can produce flux distributions within a set of possibilities, the possible range of correlation (from the minimum to the maximum) between the measured fluxes and the predicted fluxes was calculated, which is presented within square brackets above the average correlation. The calculation of these ranges is described in Appendix 3. pFBA also does not necessarily produce a unique flux distribution (as discussed in section 2.2.3.1), but calculation of the possible range of correlation is complex, and we have therefore omitted it.

In addition, we performed FBA along with the minimization of l^2 norm (Table 2.3, denoted as FBA+min l^2). It also showed good correlation with measured fluxes (0.8106, SD: 0.1740). When knowledge of uptake rates is available, the FBA+min l^2 method is a good alternative to pFBA, since it is easier to implement and produces a unique metabolic flux distribution.

2.3.3 Detailed quantitative features of the predicted fluxes

In addition to calculating correlations, we examined how the predicted and the measured metabolic flux distribution visually compare to each other. Since the predicted flux of our method has an arbitrary unit, the magnitudes of the predicted fluxes were normalized by the Euclidean norm of the measured flux vector for comparison. The results are shown in Fig 2.4. The *x*-axis represents a set of metabolic reactions used to calculate correlation between the measured and the predicted fluxes, and the *y*-axis indicates flux value. The scale and the units on the y-axis are based on those of the measured flux. The reactions are grouped functionally based on the pathways in which they are participating such as glycolysis and the tricarboxylic acid cycle. As can be seen in the figure, the predicted and the measured metabolic flux distribution look similar to each other when the correlation between them is high. We see moreover that AC+SPOT predicts negative fluxes for some reactions which are supposed to be positive, which might explain one of the reasons why the method shows relatively poor correlation compared to the other three methods (DC+E-Flux2, AC+E-Flux2, and DC+SPOT). Based on this observation, possible ways to improve the correlation of AC+SPOT will be discussed in the following section.

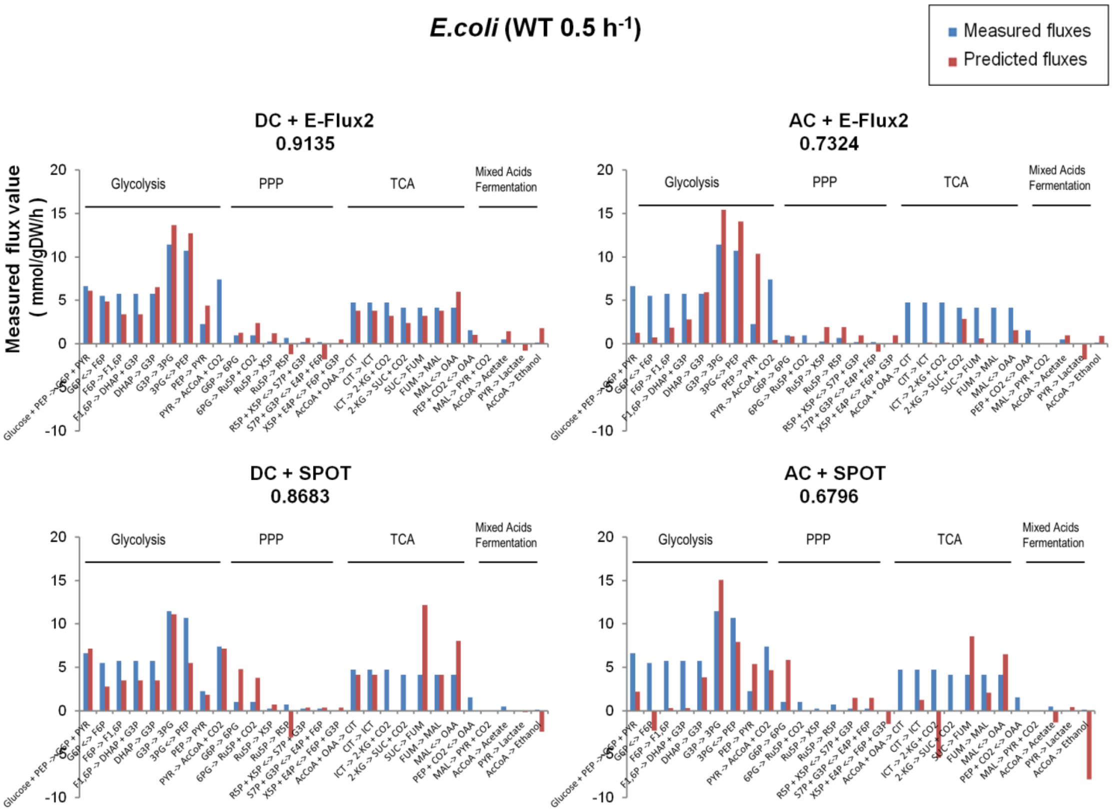


Fig 2.4 Comparison of the predicted fluxes with the measured fluxes (*E. coli*,WT 0.5h^{-1} sample). The *x*-axis represents metabolic reactions used to calculate correlation between the measured (blue bars in the figure) and the predicted fluxes (red bars in the figure), and the *y*-axis indicates flux value. The scale and the units on the *y*-axis are based on those of the measured flux.

2.3.4 Test of our methods on previous models of *E. coli* and *S. cerevisiae*

E. coli and *S. cerevisiae* are two of the most intensively studied model microorganisms. On the other hand, genome-scale metabolic models of many other organisms are still incomplete. Thus, it is important to examine the applicability of our methods to relatively incomplete models before applying our methods to other organisms. One of the ways to test this is by running our methods on older models of *E. coli* and *S. cerevisiae* that are relatively incomplete. Using the same transcriptomic and fluxomic datasets, we tested our methods on older models of *E. coli* (*i*JR904 and *i*AF1260) and of *S. cerevisiae* (*i*ND750 and *i*MM904), and the results obtained are shown in Fig 2.5. In this figure, the *x*-axis represents the four different optimization strategies (denoted as DC+E-Flux2, AC+E-Flux2, DC+SPOT and AC+SPOT) and the *y*-axis identifies the average Pearson correlation coefficient between the predicted fluxes and the measured fluxes of *E. coli* (Fig 2.5a) and *S. cerevisiae* (Fig 2.5b). Each optimization strategy consists of a group of three bars among which the blue, red, and green bars indicate the average correlation of the oldest, middle, and newest models. Error bars represent the standard error of the mean (SEM).

In case of *E. coli* (Fig 2.5a), the two recent models (*i*AF1260 and *i*JO1366) showed better average correlation than the oldest model (*i*JR904) in most cases. We found that there is little difference in the average correlation between *i*AF1260 and *i*JO1366.

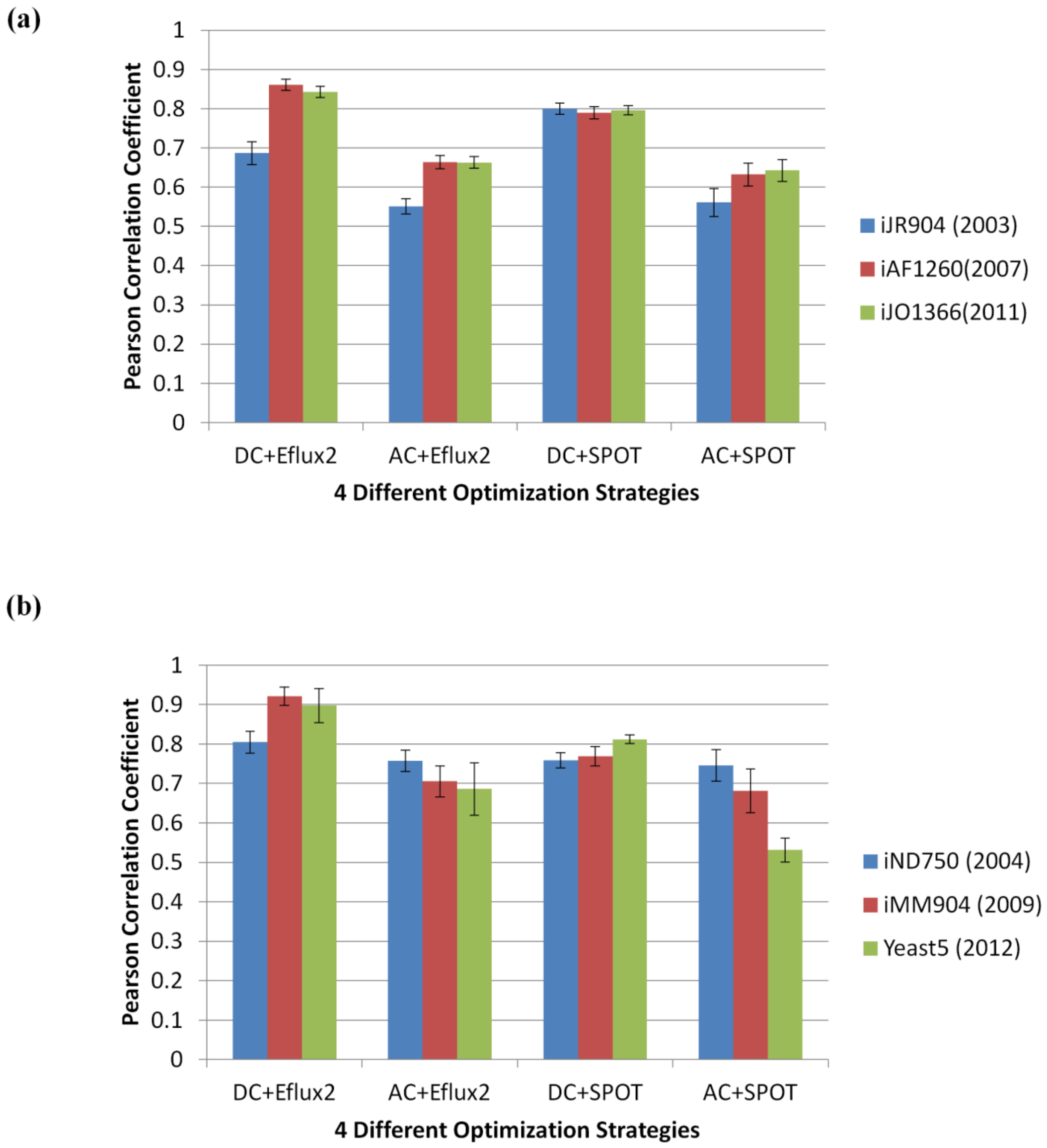


Fig 2.5 Test of our methods onto older models of *E. coli* and *S. cerevisiae*. We tested our methods on older models of *E. coli* (*i*JR904 and *i*AF1260) and those of *S. cerevisiae* (*i*ND750 and *i*MM904) to examine the applicability of our methods to the relatively incomplete models. The *x*-axis represents the four different optimization strategies and the *y*-axis identifies the average

Pearson correlation coefficient between the predicted fluxes and the measured fluxes of *E. coli* (Fig 2.5a) and *S. cerevisiae* (Fig 2.5b). Error bars represent standard error of the mean (SEM).

In the case of *S. cerevisiae* (Fig 2.5b), the newest model (Yeast 5) achieves a correlation that is essentially as good as or better than earlier models (*i*ND750 and *i*MM904) when the carbon source is known (DC+E-Flux2 and DC+SPOT). Unlike the *E. coli* case, however, the newest yeast model performs worse than the older models when carbon source of a yeast cell is unknown, especially for AC+SPOT. To understand why, we explored the hypothesis that the reason for the poor performance of Yeast5 is because of the larger number of carbon sources (Fig 2.6).

Although we could not fully identify the reason, it seems that the larger number of carbon sources has something to do with its decrease performance, but is certainly not the whole story. A different degree of interconnectivity among intracellular and exchange reactions inherent to each model or other unknown factors (e.g. thermodynamically infeasible cycles and dead-end metabolites that are unintentionally added to a newer model) may also play a role.

Except for *i*JR904 in *E. coli*, DC+E-Flux2 showed the highest average correlation (between 0.8 and 0.9) in both microorganisms. Thus, if we study an organism where information on both carbon source and objective function is known, applying DC+E-Flux2 is recommended.

Interestingly, DC+SPOT (known carbon source and unknown objective function) shows steady and constant average correlation between 0.7 and 0.8 in both *E. coli* and *S. cerevisiae* regardless of which model was used. The method seems to be the least influenced by incompleteness of the model. Thus, DC+SPOT is useful for predicting intracellular metabolic flux distribution in less well-studied organisms.

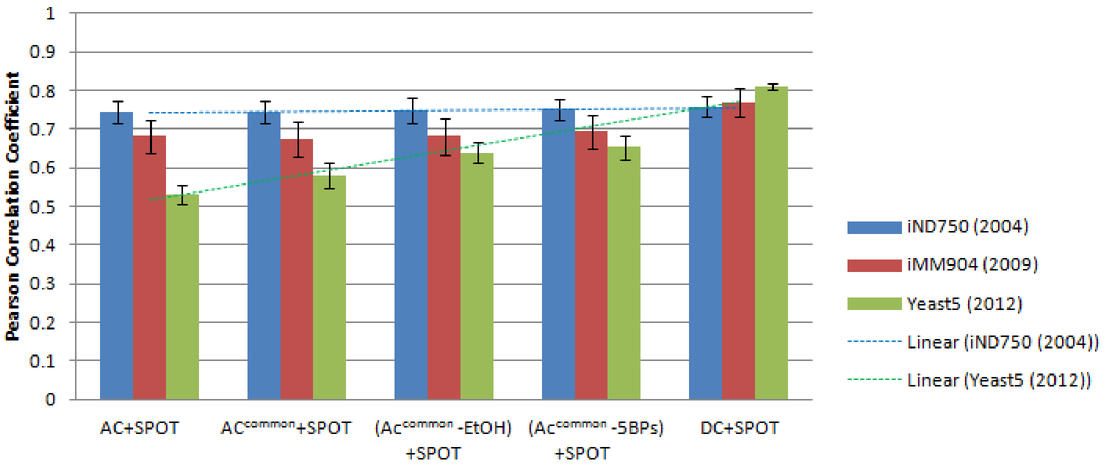


Fig 2.6 Exploration of the way to improve a poor performance of Yeast 5 in AC+SPOT. As described in section 2.2.2, we built the AC (All possible Carbon sources) model, which has lower bounds of negative infinity for all exchange reactions of possible carbon sources (i.e. external metabolites containing carbon) to simulate the unknown carbon source situation. As listed in S1 Table of Supporting Information, a total of 108, 154, and 158 potential carbon sources were allowed to be taken up by the cell for *i*ND750, *i*MM904, and Yeast 5 models of *S. cerevisiae*, respectively. To test whether Yeast 5 performs worse than the older models because it has more carbon source uptake reactions (leading to more incorrect carbon sources to confound the prediction method), we performed SPOT again after blocking the uptake of model-specific carbon sources, leaving 106 exchange reactions that are common across all three models (see S3 Table of Supporting Information for details). In other words, we updated the three yeast AC models so that they all have the same set of 106 possible carbon source exchange reactions, which we call the AC^{common} (All possible *common* Carbon sources) model in the figure to distinguish it from the original AC model.

As shown in the AC^{common}+SPOT case of the figure above, although the average correlation of Yeast 5 was improved from 0.5313 to 0.5791 after reducing the number of carbon source uptake reactions to 106, the older yeast models still outperform Yeast 5, which suggests that the lower correlation achieved by SPOT on the Yeast 5 AC model is not simply due to having a greater number of possible carbon sources.

Interestingly, unlike the older models, the performance of Yeast 5 in AC+SPOT was much improved by limiting the uptake of well-known by-products of yeast such as ethanol and glycerol [89]. As shown in the (AC^{common}-EtOH)+SPOT and (AC^{common}-5BPs)+SPOT case, the correlation of Yeast 5 was further increased from 0.5791 to 0.6397 and 0.6534, when the uptake of ethanol and of well-known five metabolic by-products of yeast (ethanol, carbon dioxide, succinate, glycerol, and acetate) was blocked respectively. The different sensitivity of the performance of each model to changes in the number of carbon sources (see the trend lines in the figure) may indicate a different degree of interconnectivity among intracellular reactions and exchange reactions inherent to the model. Considering that only a small number of preferred carbon sources are consumed by most microorganisms, which is a well-known phenomenon called Caron Catabolite Repression (CCR) [90,91], reducing the number of possible carbon source uptake reactions by analyzing growth media composition is a sensible way to improve the predictive accuracy of AC+SPOT.

2.3.5 Measurement of the speed of our methods

From a practical perspective, short running time is a desirable characteristic. Thus, we measured the running time of our algorithm for all 80 samples (4 optimization strategies and 20 samples per strategy) using the built-in MATLAB function, profile. The execution time for our method is illustrated in Fig 2.7. Regardless of which simulation strategy is used, our method, including mapping the transcriptomic data, solving the optimization problem, predicting the intracellular metabolic flux distribution and calculating the correlation with the measured fluxes, can be performed within two seconds for both microorganisms.

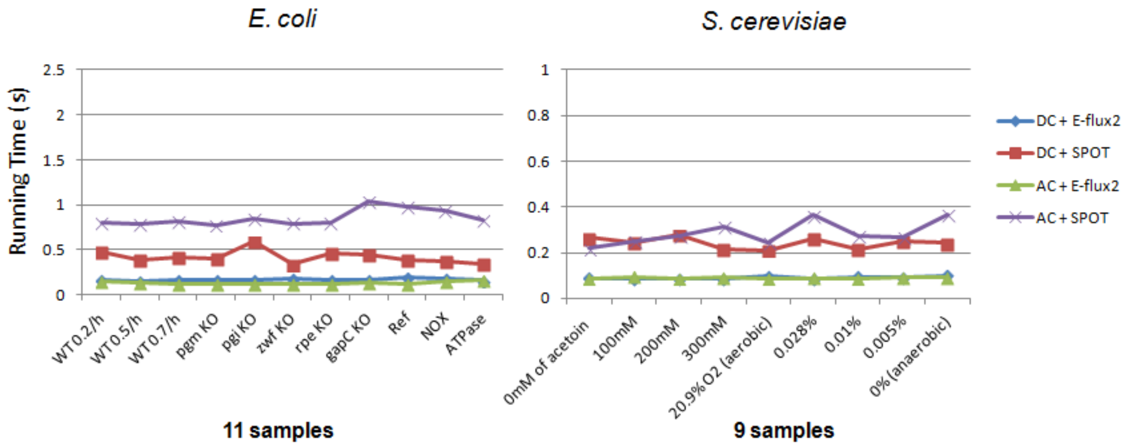


Fig 2.7 Average running time of our algorithm. We measured the running time of our algorithm implemented using MATLAB (The Mathworks, Inc., Natick, Mass.) and Gurobi Optimizer (Gurobi Optimization, Inc., Houston, Texas) for all 80 samples (4 simulation methods and 20 samples per simulation method) using the built-in MATLAB function, profile. Regardless of which simulation method is used, our method completes within one second for both *E. coli* and *S. cerevisiae*. Computations were carried out on the Window 8 OS platform using a personal computer with an Intel Core i5 3.10 GHz processor with 8 GB of RAM.

2.3.6 Implementation of our methods in a user-friendly interface

As an ultimate representation of the cellular metabolic phenotype, metabolic fluxes provide important information to understand the functioning of cellular processes [6]. Our methods which allow to quickly and easily determine metabolic fluxes from gene expression data, thus, will be of interest to a wide audience in various biological fields.

For possible users of our method especially who are not skilled in computer programming, we implemented E-Flux2 and SPOT in an intuitive user-friendly interface called MOST to make our methods viable to all researchers regardless of whether they are trained in computer science or not. MOST (Metabolic Optimization and Simulation Tool, http://most.ccib.rutgers.edu/) is an open source-based software package for constraint-based modeling [85]. It provides Excel-like editing functionality as well as supports Systems Biology Markup Language (SBML) and Comma Separated Value (CSV) files. How to run our E-Flux2 and SPOT in MOST is described in S2 File of Supporting Information.

2.4 Discussion

In this chapter, we introduced a new computational method that we developed for inferring intracellular fluxes from transcriptomic data using genome-scale models, which satisfies desirable features summarized in Table 2.1. On top of that, the predictive accuracy of our method was validated against measured intracellular fluxes, and we found it to be more accurate than existing methods.

Our method can also be easily applied to study the metabolic flux distributions of various engineered strains with little or no modification to genome-scale models since transcriptomic data themselves reflect knock-outs, knock-ins (with addition of metabolic reactions into the model that correspond to the knocked-in gene), induced-amplification or induced-repression of metabolic genes. In addition, E-Flux2 is flexible in that if there is an alternative objective to maximizing growth rate that is considered more applicable to a certain organism, then the biomass flux used in the first optimization step of E-Flux2 can be replaced with this objective function.

The multiple advantages of our method make it a useful tool for identifying fundamental mechanisms of metabolic responses and finding molecular targets for metabolic engineering. For instance, by using this tool with a set of gene expression data measured over a time course, we can determine how intracellular metabolic flux changes and where significant redirection occurs. Our method is available in a user-friendly, open source-based software package called MOST (http://most.ccib.rutgers.edu/).

CHAPTER 3: Biological applications of our computational methods

As shown in Chapter 2, E-Flux2 and SPOT, the computational methods that we developed for inferring metabolic flux distributions from transcriptomic data, overcome several shortcomings of existing methods and combine desirable characteristics including applicability to a wide range of experimental conditions, production of a unique solution, fast running time, and the availability of a user-friendly implementation (at http://most.ccib.rutgers.edu/). Most importantly, the predictive accuracy of our method was validated using the largest experimental dataset compiled to date, consisting of 20 experimental conditions of gene expression measurements coupled with corresponding central carbon metabolic flux measurements (11 in *E. coli* and 9 in yeast). Our method outperformed a representative sample of competing methods in terms of the average of correlations between predicted and measured fluxes. The models, codes and dataset used for this study are publicly available so that other researchers can also use them.

The goal of developing these computational tools is to better understand complex biological systems. Not only do the methods we developed contribute to advancing previous work, they have helped to answer biological research questions as well, as can be seen in several collaborative publications. In these publications, our methods were used to understand the lipid accumulation mechanism of nitrogen-stressed *Phaeodactylum tricornutum* cells [92], verify the predictive power of a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002 [93], and examine the metabolic impacts of RpiRc, a potent repressor of microbial toxins in *S. aureus* [94].

3.1 Application 1: Understanding the lipid accumulation mechanism of *P. tricornutum* under nitrogen starvation

3.1.1 Background

Biofuels produced from sunlight, carbon dioxide and water by photosynthetic microorganisms are one of the promising sustainable energy sources that can displace petroleum-derived fuels [95]. Diatoms are a diverse group of eukaryotic, unicellular, and photosynthetic microalgae that are responsible for at least a quarter of inorganic carbon fixed each year in the ocean [96]. The marine diatom, *Phaeodactylum tricornutum*, is classified in the phylum Bacillariophyta, and this phylum comprises one-third of all known marine phytoplankton. *P. tricornutum* has been extensively studied as a model diatom in the context of physiology, biochemistry, and genomics [97].

One of the interesting features of diatoms including *P. tricornutum* is that they accumulate significant amount of storage lipids, mainly in the form of triacylglycerols (TAGs) when nitrogen availability decreases and limits their growth (Fig 3.1) [92]. Since TAG is a direct precursor of biodiesel, understanding how carbon flow is pushed into TAG synthesis has been a topic of interest for many researchers who seek for a strategy for high-yield production of algal-based biofuels [98,99]. Using our methods, we examined how a diatom remodels intermediate metabolism to respond to nitrogen stress.

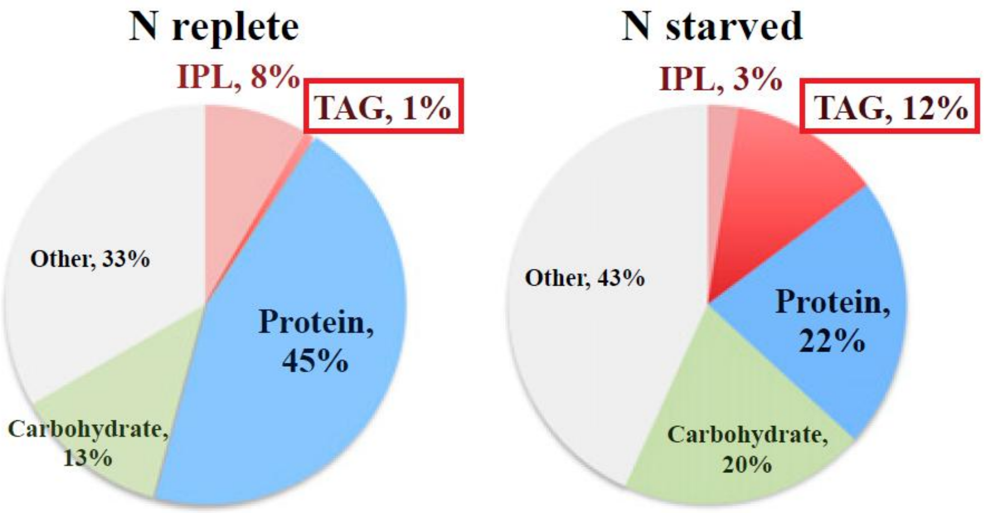


Fig 3.1. Allocation of cellular carbon to different biosynthetic compounds in nitrogen-replete and -stressed *P. tricornutum* 48 h after inoculation

3.1.2 Materials and Methods

3.1.2.1 Cultivation and experimental planning

P. tricornutum was obtained from the Provasoli–Guillard National Center for Culture of Marine Phytoplankton (accession Pt1 8.6). It was maintained axenically in sterile artificial seawater supplemented with F/2 nutrients and buffered with 2 mmol/L Tris to pH 8 [100–102]. Pre-inocula were grown with NaNO3 (0.88 mmol/L) as the sole nitrogen source. After 48 h of growth, exponentially growing pre-inocula were centrifuged (5500 g, 10 min) and washed twice with nitrate-free F/2 medium and inoculated in triplicate at a concentration of 2.5 × 105 cells/mL into fresh F/2 medium with nitrate (0.88 mmol/L, nitrogen-replete condition), and F/2 medium without any nitrogen source (nitrogen-free condition). Cell densities were determined using a

Beckman Coulter Multisizer 3 (Beckman Coulter Inc.). The set of three optically thin, biologically independent cultures was maintained under exponential growth conditions in UltraCruz flasks at 18 °C and 120–150 μmol photons/m^2·s continuous white light emitting diodes (LEDs) and aerated through 0.2-μm filters. Both treatments were sampled after 48 h to assure the largest contrast between the physiological states. Cultures were sampled while the N-replete culture was in exponential growth and the N-stressed culture stopped dividing for 24 h; however, both cultures were still optically thin. Known numbers of cells were then filtered, flash frozen in liquid nitrogen, and kept at 80 °C until analysis.

3.1.2.2 Genome-scale metabolic network of *P. tricornutum*

For this study, we used the genome-scale stoichiometric metabolic model of *P. tricornutum* constructed by Kim et al. [103]. It was built using genomic databases, DiatomCyc (http://www.diatomcyc.org/) [104] and journal articles. The reconstruction started using annotations from the KEGG database (http://www.genome.jp/kegg/) [105] (Kanehisa and Goto, 2000) and DiatomCyc. The network includes glycolysis, the TCA cycle, oxidative and reductive pentose phosphate pathways, the phosphoketolase pathway, and amino acid, chlorophyll, nucleotide, chrysolaminarin and lipid synthesis pathways. The reversibility of reactions was then assessed using a combination of databases such as DiatomCyc, KEGG, MetaCyc [106] and BRENDA [107]. In general, there was good agreement among all three databases in terms of the reversibility of reactions. However, there were some discrepancies between the databases; in these cases, BRENDA was used to determine reaction reversibility. Total 34 of ‘gap’ reactions that were not found in databases were added if the majority of the enzymes in a network were present, but the network was incomplete without including these missing reactions. This process was done based on the assumption that the alga possesses the complete pathway, but the enzymes were missed in annotation. The gaps in the network were addressed by first searching for the missing enzymes in other organisms, then using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [108] to search the *P. tricornutum* database for similar amino acid sequences. The metabolic network of *P. tricornutum* consists of 587 metabolites and 850 metabolic reactions, including the biomass equation.

3.1.2.3 Transcriptome measurements by RNA-Seq

Samples from both the nitrogen-replete and -stressed cultures for RNA-Seq were harvested by centrifuging 5×10^7 cells for 8 min at $6{,}000 \times g$ using a Sorvall RC6+ Centrifuge (Thermo Scientific) at 4 ℃. The samples were frozen in liquid N_2 and stored at −80 °C. Total RNA was extracted using an RNAeasy Plant Mini Kit (Qiagen) followed by removal of DNA contamination using Ambion Turbo DNase (AM1907; Life Technologies). PCR was performed to confirm that there was no DNA contamination. Total RNA quantification and quality were assessed spectrophotometrically with a Nanodrop 1000 (Thermo Scientific). TruSeq RNA (Illumina) was used to prepare mRNA libraries for each sample conforming to the manufacturer's instructions. The 50-bp single-ended libraries of all samples were multiplexed (pooled) and sequenced on an Illumina MiSeq platform. After being trimmed for adaptor and low-quality sequences, the raw reads were then aligned to *P. tricornutum*'s version 2.0 set of 10,402 filtered gene models (http://genome.jgi.doe.gov/Phatr2/Phatr2.info.html) using CLC Genomics

Workbench (v6.02) [109]. Functional metabolic assignment for the different gene models were done according to KEGG, DiatomCyc, and published literature [110,111]. After aligning the raw data to the *P. tricornutum*'s gene models, files were filtered to retrieve uniquely aligned reads with no more than three mismatches. Gene counts (unique aligned reads per gene) were used for DE analysis carried out using the DESeq R/Bioconductor package [112], which infers differentially expressed genes (DE) based on the negative binomial distribution. For this analysis, we used a cutoff of 5% to control for false positives and considered only genes that had a log twofold change greater than or equal to ±2 and a false detection rate < 0.05 to be DE. The output of DESeq for all 10,402 genes was submitted to the Gene Expression Omnibus (GEO) under accession no. GSE56346 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56346), and all reads were deposited to the National Center for Biotechnology Information's Short Read Archive under accession no. SRP040703 (https://www.ncbi.nlm.nih.gov/sra/SRP040703).

3.1.2.4 Metabolic flux prediction

Computational prediction of metabolic fluxes was performed using a prototype of SPOT (see Chapter 2, section 2.2.3.2) [113]. That is, based on a given limited translational efficiency and a limited accumulation of enzyme over the time, we set the objective function to maximize the correlation between the flux vector and a vector of corresponding transcriptomic data. .To calculate this correlation, we used the uncentered Pearson product–moment correlation, a measure of the linear correlation between two vectors:

$$\max \ \frac{v_{\mathrm{irr}} \cdot g_{\mathrm{irr}}}{\|v_{\mathrm{irr}}\| \|g_{\mathrm{irr}}\|}$$

$$\text{subject to}\begin{cases} Sv = 0 \\ a_j \leq v_j \leq b_j \end{cases}$$

where v_{irr} is a flux vector which represents the reaction rates of the irreversible reactions in the network; g_{irr} is a vector that indicates corresponding gene expression data; v is a flux vector representing the reaction rates of the n reactions in the network; S is the stoichiometric matrix; and a_j and b_j are the minimum and maximum reaction rates through reaction j. When maximizing the correlation, we used the set of irreversible reactions, since the directions of reversible reactions are undefined, whereas gene expression data values are always positive. In this analysis, the minimum and maximum reaction rates of the flux balance analysis were set based on the expression level of the genes associated with each reaction. The predicted fluxes were then normalized by growth rates measured under the N-replete and N-depleted conditions with an assumption of steady-state growth [100,114]. The metabolic fluxes layout was generated by Cytoscape v.2.8.3 [115].

3.1.3 Results

We calculated predicted fluxes from the transcriptomic data measured under N-replete and N-starved conditions to examine differences in the metabolic state of *P. tricornutum* between the two conditions. The analysis showed that, in general, most metabolic reactions were more active under the N-abundant condition. To be more specific, 92% of the predicted fluxes were smaller in the N-starved condition than in the N-replete condition, which makes sense since nitrogen-stressed *P. tricornutum* cells tend to show a lower growth rate and a lower photon turnover rate. However, there were some exceptional cases. First, as shown in Fig 3.2, predicted fluxes of the nitrogen assimilation pathway, which is known to produce amino acids from inorganic nitrogen sources, were significantly higher under the N-starved condition. The activated metabolic reactions include those that are mediated by three key enzymes in nitrogen assimilation: Glutamine synthetase (GS) which incorporates ammonia into glutamate to form glutamine; Glutamine 2-oxoglutarate aminotransferase (GOGAT, also known as glutamate synthase) that converts 2-oxoglutarate (also called α-ketoglutarate) and glutamine to two molecules of glutamates, which works together with GS to replenish glutamate so that the GS reaction is not substrate-limited; and Glutamate dehydrogenase (GDH) which interconverts glutamate and 2-oxoglutarate utilizing either NAD(P)+ or NAD(P)H.

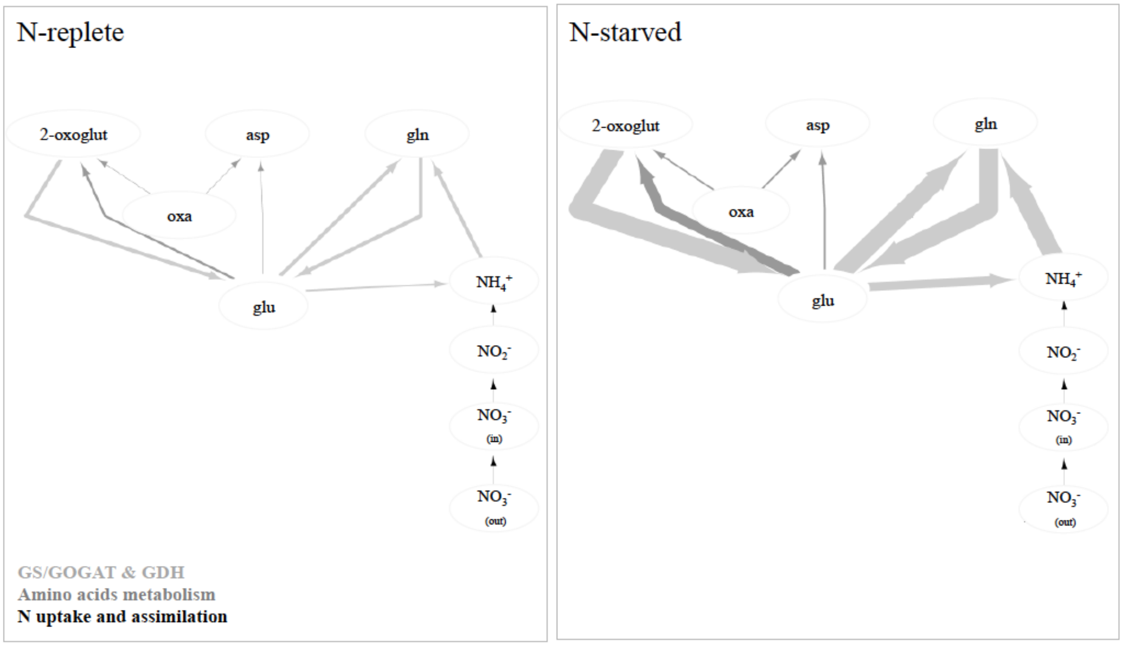


Fig 3.2. Predicted metabolic fluxes of nitrogen metabolic pathways under nitrogen-replete and -stressed conditions. The width of each arrow represents the relative volume of the flux as calculated by the model. Under N-starved condition, metabolic reactions related to the

assimilation of internal nitrogen sources were more active. Dark gray, glutamate dehydrogenase- and aspartate aminotransferase-related fluxes; light gray, GS/GOGAT pathway; black, nitrogen uptake and assimilation related fluxes. Abbreviations: arg, arginine; asp, aspartate; GDH, glutamate dehydrogenase; gln, glutamine; glu, glutamate; GS, glutamine synthetase; NH_4^+, ammonium; NO_2^-, nitrite; NO_3^-, nitrate; oxa, oxaloacetate; 2-oxoglut, 2-oxoglutarate.

Active nitrogen assimilation under nitrogen-starved condition sounds contradictory. However, the second exceptional example of the active metabolic reactions under N-stressed condition provides a possible explanation for where the inorganic nitrogen sources come from. In addition to the nitrogen assimilation pathway, predicted fluxes of the urea and TCA cycles also increased significantly (Fig. 3.3). One of the functions of the urea cycle is to recycle degradated proteins by regenerating the ammonia groups from their carbon backbones. Those recycled inorganic nitrogen sources can be reused for nitrogen assimilation [116]. The remnant carbon skeletons can also get reprocessed as intermediate metabolites of the TCA cycle which provides precursors for anabolic processes including lipid biosynthesis and reducing factors (i.e. NADH and $FADH_2$) that drive the generation of energy [117].

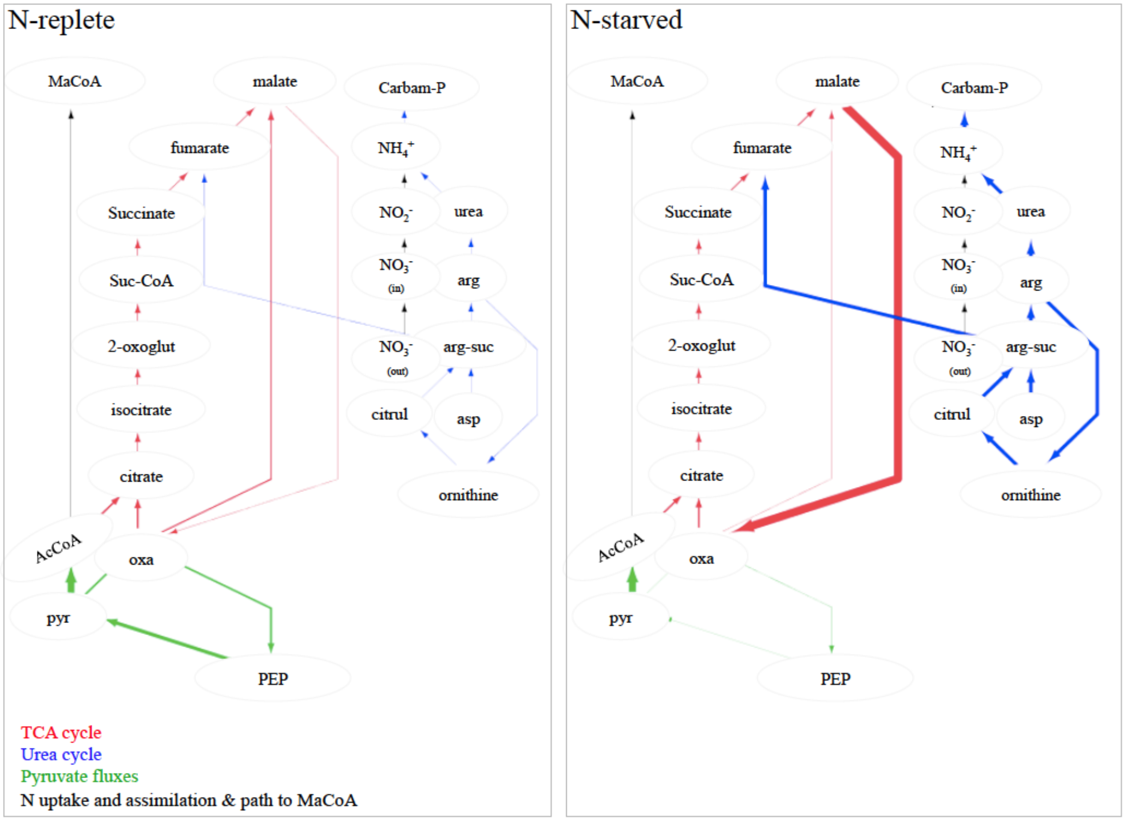


Fig 3.3. Predicted metabolic fluxes of intermediate carbon metabolic pathways under nitrogen-replete and -stressed conditions. The width of each arrow represents the relative volume of the flux as calculated by the model. Under N-starved condition, metabolic reactions related to the TCA and the urea cycles were more active. Fluxes are color-coded according to their pathway: black, nitrogen uptake and assimilation and shunting carbon toward FA metabolism; blue, urea cycle; green, pyruvate-related; red, TCA cycle. Abbreviations: acCoA, acetyl-CoA; arg, arginine; arg-suc, argininosuccinate; asp, aspartate; carbam-P, carbamoyl-phosphate; citrul, citrulline; maCoA, malonyl CoA; NH_4^+, ammonium; NO_2^-, nitrite; NO_3^-, nitrate; oxa, oxaloacetate; 2-oxoglut, 2-oxoglutarate; PEP, phosphoenolpyruvate; pyr, pyruvate; suc-CoA, succinyl CoA.

All in all, this computational analysis suggests that nitrogen stress leads to a remodeling of intermediate metabolism centered around glutamate, one of the primary nitrogen carriers in most organisms. This metabolic hub involves transfer of an amino group to form glutamine and subsequent reactions involving both GDH and GOGAT (Fig. 3.2), and seems to conserve intracellular nitrogen level by nitrogen assimilation from inorganic nitrogen sources derived from protein degradation products recycled through the urea cycle (Fig. 3.3). Our metabolic flux analysis further predicts that this hub is coupled to the intermediate metabolism of carbon in the TCA cycle. Specifically, the analysis predicts an increase in the oxidation of malate to oxaloacetate, resulting in an increased catabolic production of NADH which is the preferred biological reducing agent for fatty acid synthesis (Fig. 3.2). In contrast, the central hub that involves pyruvate, phosphoenolpyruvate, and oxaloacetate is down-regulated in nitrogen-stressed cells, and this intermediate metabolic pathway is unlikely to be a significant source of carbon for lipid biosynthesis.

3.1.4 Discussion

Under nitrogen stress, *P. tricornutum* becomes chlorotic, and cannibalizes and remobilizes its plastid protein towards energy storage, mainly in the form of TAGs (Fig 3.1). Like in all other eukaryotic microalgae, the intermediate metabolism of carbon and nitrogen metabolism is closely coupled with a hub centered around glutamate and 2-oxoglutarate in diatoms. Our results suggest that this hub operates to redirect intracellular nitrogen derived from the catabolism of enzymes that are temporarily incapable of supporting growth toward a few selective enzymes that confers advantages under nitrogen stress. The selected enzymes are primarily associated with degradation of proteins (e.g. the urea cycle), nitrogen assimilation (e.g. GS, GOGAT, and GDH), and energy-generating/anaplerotic reactions such as the TCA cycle (Figs. 3.2 and 3.3). This observation is in line with other previous studies with wild-type *P. tricornutum* reporting the up-regulation of genes in the TCA cycle and the nitrogen assimilation pathways under nitrogen stress [97,118].

Our results strongly suggest that the remodeling of intermediate metabolism under nitrogen stress is the result of at least two different processes (Fig. 3.4). The first process is related to shuffling of preexisting protein nitrogen as reflected by the loss of photosynthetic machinery. The second process is associated with redirecting photosynthetically fixed carbon from amino acids to other sinks, especially lipids. Although nitrogen-stressed cells recycle the nitrogen from preformed photosynthetically fixed carbon toward nitrogen-deficient storage molecules, they continuously synthesize carbon skeletons, mainly in the forms of 2-oxoglutarate, fumarate, and malate (Figs. 3.2 and 3.3). These relatively oxidized intermediate metabolites become potential sinks for photosynthetically produced reductants, leading to the formation of lipids (Fig. 3.3). In a subsequent study [119], our collaborators were able to enhance lipid biosynthesis in *P. tricornutum* without causing a significant impairment in its photosynthetic capacity by knocking down nitrate reductase (NR) which is a key enzyme in nitrogen assimilation; this allows the cells to utilize most of the carbon skeletons that are derived from degradated proteins for lipid biosynthesis. Overall, our method was used to explain the underlying mechanisms responsible for the metabolic responses of *P. tricornutum* under nitrogen-starved condition. In addition, the computational analysis results helped to generate a hypothesis for a subsequent study and to provide a relatively simple metabolic engineering strategy (i.e. single gene knock-down) for redirecting carbon sources of the model diatom toward lipid production.

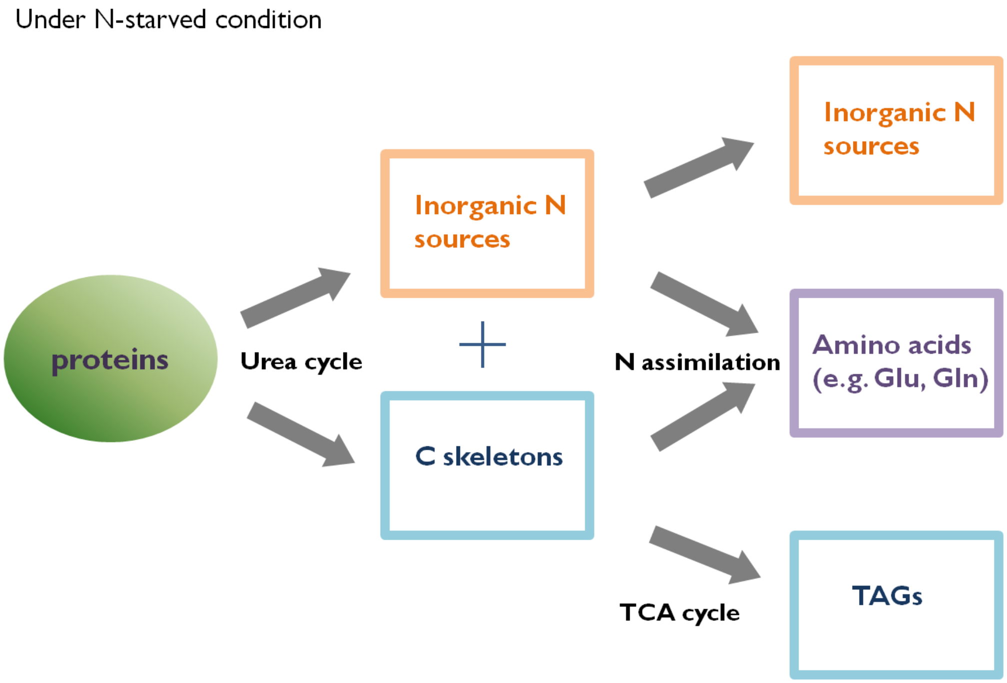


Fig. 3.4. Schematic representation of the mechanism suggested by our method on the remodeling of intermediate metabolism in nitrogen-stressed *P. tricornutum* cells.

3.2 Application 2: Validation of the predictive power of a genome-scale metabolic model of *Synechococcus* sp. PCC 7002

3.2.1 Background

Cyanobacteria and microalgae are examples of unicellular aquatic microbial oxygenic photoautotrophs, most of which are recognized as the most efficient photosynthetic organisms at converting carbon dioxide, water and nutrients into complex molecules and biomass using solar energy [120,121]. For this reason, they are being pursued as leading candidates for producing chemical and biofuel precursors. Key attributes of these organisms that are important for these applications include biomass composition (lipids, carbohydrates, pigments, and proteins), nutrient and solar input requirements for growth, tolerance to environmental stress, and ease of genetic transformation. As these attributes differ greatly among species, there is considerable interest in understanding species variations in metabolism and its regulation under environmental stimuli.

Genetic modification is a successful strategy for redirecting metabolic intermediates into chemicals of biotechnological interest. Removing competing pathways, overexpression of native pathways or addition of new pathways have each been applied to elevate the yield of desired biochemicals [122–124]. However, simple transgenic mutants often exhibit undesired attributes such as slower growth rate, susceptibility to environmental stresses, and production of unanticipated products because of modified regulation of metabolic pathways and misbalance of energy resources [125,126]. Researchers have explored more complex genetic changes involving replacement of whole metabolic pathways and transcriptional regulation of multiple gene targets [127]. These complex strategies are time consuming and costly, and are best implemented together with a computational model of metabolism capable of simulating genetic modifications.

Genome-scale network reconstructions of metabolism are built from all annotated metabolic genes and corresponding reactions in an organism of interest. A network reconstruction can be converted into a mathematical format, and thus be used for mathematical and computational analysis. On the qualitative side of the genome model, its predictions allow testing gene essentiality and effects of environmental perturbations much more quickly. On the quantitative side, model predictions can decipher ratios of nutrient utilization, central carbon metabolism fluxes, cell growth and nutrient exchanges under different growth conditions. Quantitative phenotype predictions have proven to be particularly useful capability for bioengineering applications.

So far, genome-scale models have been constructed for more than 100 organisms [128]. Among these organisms, *Synechocystis* sp. PCC 6803, a freshwater strain, is the most well modeled and extensively studied cyanobacterium [129,130]. In addition, genome-scale models have also been constructed for a marine cyanobacterial strain, *Synechococcus* sp.PCC 7002 [131–133], which has a fast growth rate (0.20 h^{-1}), well studied genomic database and mature transformation protocol [134,135].

In the present study we constructed a metabolic model of *Synechococcus* sp.PCC 7002, *i*Syp821, newly incorporating a variable biomass objective function in which stoichiometries of the major biomass components vary according to light intensity, which reflects fundamental property of all photoautotrophically growing microorganisms [136,137]. Additionally, *i*Syp821 was modified to account for changes in gene products (enzymes) from experimentally measured transcriptomic data and applied to estimate changes in metabolic flux distributions arising from nutrient stress. Using this strategy, we found that *i*Syp821 correctly predicts the observed redistribution pattern of carbon products under nitrogen depletion, including decreased rates of CO_2 uptake, amino acid synthesis, and increased rates of glycogen and lipid synthesis.

3.2.2 Materials and Methods

3.2.2.1 Metabolic network reconstruction

A web-based tool denoted Model SEED (available at http://www.theseed.org/models/) was used to input the map of metabolic reactions based on available curated genomic data. The model was curated according to the step-wise procedure described in [138]. Additionally, the curation process of the model was enhanced by following the most recently published genome-scale network reconstruction rules [139–141]. A summary of the features defining the *i*Syp821 model are given in Table 3.1 below.

The reconstruction of *i*Syp821, references for the biochemical reactions, and the list of genes and reactions newly added are given in Data S1, Data S2, and Data S4 at http://www.sciencedirect.com/science/article/pii/S0005272816306909. *i*Syp821 includes all the reactions and genes in the two previously published *Synechococcus* sp.PCC 7002 models. A comprehensive literature review was performed to assign accurate confidence scores and proper references to all the reactions in the network. MOST (http://most.ccib.rutgers.edu/) [85] was used to correct errors in the draft model, arriving at our final reaction network. A schematic overview of model construction, curation, and database usage is provided in Fig. 3.5. The photosynthetic electron transfer chain was manually added into the metabolic network after the automated procedures were finished.

		*i*Syp821 (This paper)
Genes		821
Reactions (Metabolic and transport)	GPR	723
	NGPR	21
Exchange reactions		48
Metabolites		777

Table 3. 1 Feature summary of *i*Syp821

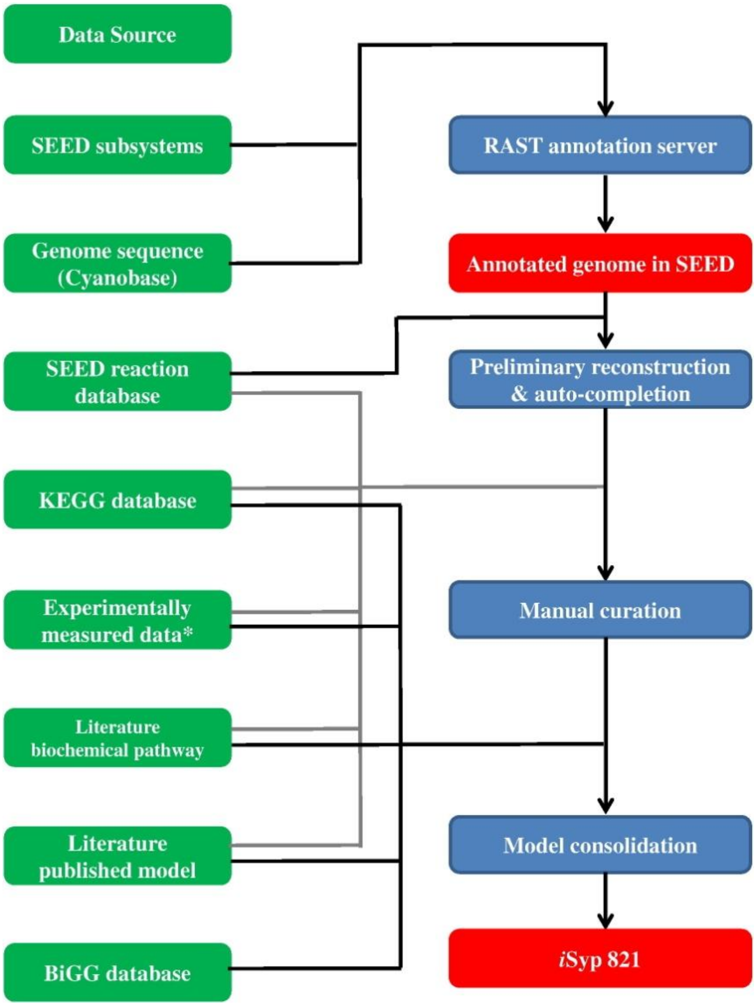


Fig. 3.5. Pipelines to construct *i*Syp821. Database and experimental data are colored in green, major construction steps are colored in blue, and versions of model are colored in red.

3.2.2.2 Transcriptome-constrained carbon metabolic fluxes under nitrogen -deprivation

Two growth conditions were used with this method: nitrate-replete and nitrate-starved. Nitrate (NO_3^-) was the sole external nitrogen source in all experiments. The mathematical constraints and strategy for integration of transcriptomic data in genome-scale metabolic model using E-Flux2 (see Chapter 2, section 2.2.3.1) [113]. Experimentally measured transcriptomic data were obtained from a previous study by Ludwig and Bryant [142], which were integrated into the genome-scale metabolic model according to the gene-protein-reaction (GPR) associations in *i*Syp821. Biomass production was taken as the objective function, and the biomass equations used for this study can be found in Data S3 at http://www.sciencedirect.com/science/article/pii/S0005272816306909. The predicted flux was normalized by growth rates experimentally measured under different nitrogen conditions (0.19 h^{-1} with nitrate, 0.08 h^{-1} without nitrate).

3.2.3 Results

Nitrogen removal from growth media is widely used to stimulate redistribution of stored carbon from proteins into carbohydrates in cyanobacteria and into lipids in microalgae [143–145]. We adapted *i*Syp821 to model these changes by integrating transcriptomic data into the model to account for changes in mRNA levels. The gene expression data can be found in Data S6 at the following web address:

http://www.sciencedirect.com/science/article/pii/S0005272816306909. (collected from Ludwig & Bryant [146]). The carbon flux distribution of photoautotrophic cultures grown under nitrogen-deprived condition was then simulated using E-Flux2 [113] and compared to nitrate-replete growth condition (Fig. 3.6). Under simulated nitrogen-deprived conditions, the model predicts these flux changes: 1) the absolute fluxes of most of the pathways in the map, except those into glycogen and Malonyl-CoA (lipid precursor) (Fig. 3.6A) were predicted to decrease by 30% or more, 2) While the relative fluxes going into glycogen synthesis and lipid synthesis were predicted to increase more than 2-fold (Fig. 3.6B), and the relative fluxes going into the lower glycolytic pathway and into the TCA cycle were predicted to decrease more than 30% (Fig. 3.6B).

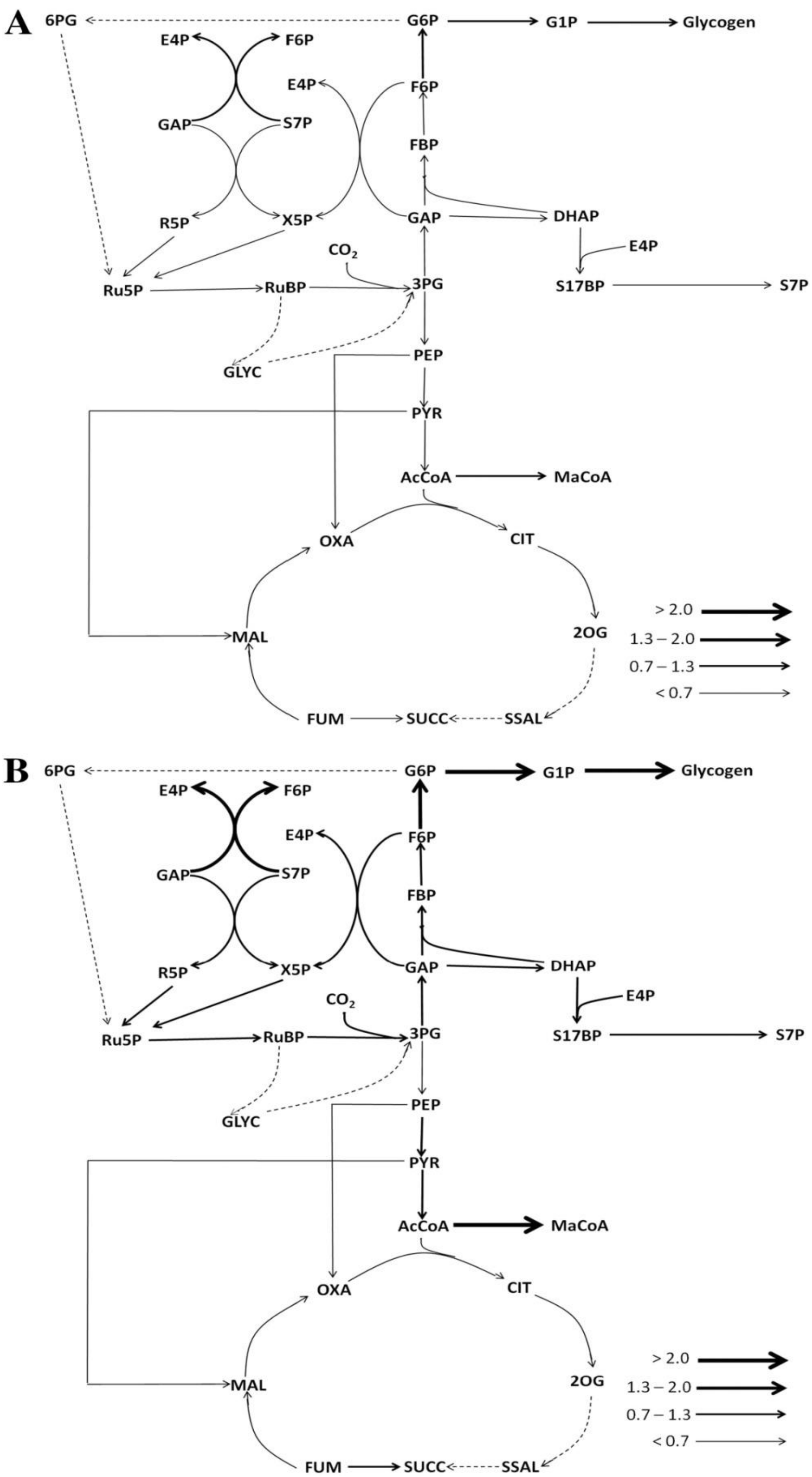


Fig. 3.6. Carbon flux redistribution under nitrogen deprivation photoautotrophic conditions as predicted by transcriptome-integrated *i*Syp821 and E-Flux2. A) Absolute carbon flux

distribution, and B) relative carbon flux distribution normalized to net CO_2 uptake. The increasing thickness of the lines represents the fold-increase of the reaction fluxes under nitrogen-starved versus nitrate-replete condition. Dashed lines indicate carbon fluxes going through these pathways are insignificant under nitrogen deprivation. Abbreviations: GAP, glyceraldehyde-3-phosphate; S17BP, Sedoheptulose 1, 7-bisphosphate; S7P, Sedoheptulose-7-phosphate; FBP, Fructose 1, 6-bisphosphate; F6P, Fructose-6-phosphate; DHAP, Dihydroxyacetone phosphate; PEP, Phosphoenolpyruvate; 2OG, Alpha-ketoglutarate; SSAL, Succinyl-semialdehyde; SUCC, Succinate; 2PG, 2-phosphoglycolate; GOL, glycolate; GOX, glyoxylate; SER, L-serine; GLY, glycine; HPYR, hydroxypyruvate.

All of these observations were validated by existing literature or experimental data. To be more specific, it is well known that during nitrogen deprivation, pigment catabolism (chlorosis) and protein catabolism occur in cyanobacteria as a means to recycle nitrogen for survival [147,148]. Our prediction of greater than 30% reduction in relative carbon flux going into both lower glycolysis pathway and TCA cycle agrees well with a lowered protein synthesis rate under nitrogen deprivation. These two pathways are the major sources of carbon precursors for protein synthesis. Associated with the loss of photosynthetic pigments and phycobilisomes, the photosynthetic activity is drastically reduced. In the cyanobacterium *Synechococcus* sp. PCC 7942, the photosystem II (PSII) and photosystem I (PSI) activities of nitrogen deprived cells retained only 0.1% of the activity of actively growing cells [147]. In this study, the PSII electron transport rate (ETR) of *Synechococcus* 7002 was predicted to be reduced by 64% under the nitrogen deprived condition where we grew the culture. Additionally, it has been shown that

RuBisCO content is depleted during nitrogen-deprived photoautotrophic conditions in *Synechocystis* 6803 [149]. Reduced PSII and PSI activities along with reduced RuBisCO abundance under nitrogen deprivation will lower the CO_2 fixation rate, in accord with our flux prediction of reduced absolute CO_2 fixation rate. Triacylglycerides (neutral lipids) are not typically stored in most cyanobacteria, while polar lipids comprise the main components of membranes. Although minor components of total biomass, lipid content has been shown to increase in several cyanobacterial species under nitrogen deprivation and photoautotrophic growth [150,151]. The transcriptomic data-integrated *i*Syp821 predicted that relative flux into malonyl-CoA, the precursor of lipid synthesis, increased by 2-fold under nitrogen deprivation (Fig. 3.6B). In most cyanobacteria, glycogen rather than protein becomes the dominant carbon sink under nitrogen-deprived photoautotrophic conditions [145]. It has been reported that nitrogen starvation during 24 h of photoautotrophic growth with continuous light increases the total carbohydrate content by 5-fold (mainly glycogen) in *Synechococcus* 7002 [152]. In line with this observation, our model predicted that the flux into glycogen synthesis increased by more than 2-fold under nitrogen deprived growth (Fig. 3.6).

3.2.4 Discussion

*i*Syp821 was constructed for simulating photoautotrophic growth of cyanobacterium *Synechococcus* 7002. One of the important new features of *i*Syp821 is that it incorporated a light-dependent biomass objective function, which allows reliable simulations of photoautotrophic growth at different light intensities. The predictions of this model were validated experimentally by data from biomass growth rate, inorganic carbon (e.g. HCO_3^-

and CO_2) uptake rate and the carbohydrate/protein content as a function of light intensity [93]. The model reveals the important transition with increasing light intensity: The relative flux ratio, PSI/PSII decreases as fluxes through both PSI and PSII increase. Additionally, the model predicts: 1) an unconventional gluconeogenesis-PP pathway that converts fixed CO_2 into carbohydrates under light, and 2) oxygenation activity of RuBisCO (photorespiration) is about 2% of its carboxylation activity at the light intensities simulated. Flux through this hybrid pathway and photorespiration activity were experimentally verified by kinetic ^{13}C metabolite labeling experiments. By incorporation of transcriptomic data to approximate enzyme concentration changes, we extended *i*Syp821 to simulate photoautotrophic carbon flux redistribution under nitrogen stress conditions. We obtained quantitative agreement with experimental data in accumulation of carbohydrates and lipids, loss of CO_2 uptake, and amino acid synthesis.

3.3 Application 3: Examination of metabolic roles of RpiRc, a potent repressor of leukocidins, in *S. aureus*

3.3.1 Background

Staphylococcus aureus is a daunting human pathogen that causes a range of diseases, from mild skin and soft tissue infections to debilitating and life-threatening bacteremia. To establish a successful infection, *S. aureus* secretes a variety of immunomodulatory proteins and virulence factors, a substantial number of which target leukocytes [153,154]. A complex family of these secreted proteins is leukocidins [155,156]. These toxins consist of two different subunits that are secreted as water-soluble monomers. The binding subunit anchors to leukocytes through host receptors, recruits the other subunit, oligomerizes, and subsequently forms pores within the host plasma membrane, leading to cell death [155,156].

The success of *S. aureus* as a versatile pathogen relies in part on its ability to infect nearly all sites of the body. This adaptability depends on the ability of *S. aureus* to fine-tune the production of its virulence factors by sensing and responding to a diversity of external stimuli, leading to optimal pathogenesis, depending on the environment it is inhabiting [157–159]. Therefore, identifying and characterizing regulators of toxins that sense and respond to disparate environmental conditions may shed light on *S. aureus* pathogenesis, enabling better therapeutic approaches for *S. aureus* clearance from specific infection sites.

The regulatory mechanisms governing the expression of leukocidins are incompletely defined. Rot (repressor of toxins) [160], is one of the well-known leukocidin repressors, which directly targets the leukocidin promotor [161,162]. However, considering the pathogen's versatile adaptability, it is more likely that there are multiple and fine-tuned mechanisms for regulating the level of leukocidins. In this study, we identified a new transcriptional regulator of leukocidins in USA300, which is the leading cause of the current community-associated methicillin-resistant *S. aureus* epidemic in the United States (here referred to as USA300). We demonstrated that inactivation of a metabolic regulatory gene, *rpiRc*, increases *S. aureus* cytotoxicity for human neutrophils; RpiRc is a potent repressor of leukocidins. Since RpiRc traditionally has been known to belong to a family of transcriptional regulators with roles in the regulation of enzymes involved in sugar catabolism in many bacterial species, the effects of inactivation of *rpiRc* on the metabolic network in USA300 was also examined using our method.

3.3.2 Materials and Methods

3.3.2.1 Culture conditions and bacterial strains

S. aureus strains were grown at 37℃ on tryptic soy agar (TSA) or in TSB with antibiotic supplementation. *E. coli* DH5α was used for cloning and propagation of plasmids. *E. coli* bacteria were grown in Luria-Bertani broth with appropriate antibiotics. Liquid cultures were grown in 5 ml of growth medium in 15-ml tubes incubated at a 45° angle with shaking at 180 rpm. For all experiments involving the growth of *S. aureus* bacteria, a 1:100 dilution of overnight cultures was subcultured into fresh medium. All the strains, plasmids, and oligonucleotides used in this study can be found in Table st5 at http://mbio.asm.org/content/suppl/2016/06/08/mBio.00818-16.DCSupplemental. The LAC *rpiRc::bursa* mutant strain was generated by phage transduction of the JE2 *rpiRc::bursa* (NE1142) strain of the Nebraska Transposon Mutant Library with phage φ80 into wild-type, erythromycin-sensitive LAC clone AH1263 [163]. Complementation of *rpiRc* on the chromosome was performed with suicide plasmid pJC1306 (kindly provided by John Chen), which is used to stably integrate DNA into the SaP1 site, resulting in a single-copy chromosomal insertion [164].

3.3.2.2 RNA isolation, RNA-Seq, and data analyses

Sample preparation for RNA sequencing was performed as previously reported by Carroll et al. [165]. Briefly, RNA was isolated using the RNeasy kit (Qiagen) and DNA was depleted employing the TURBO DNA-free kit (Ambion). The successful depletion of DNA was verified via PCR and the quality and concentration of the RNA evaluated using the Agilent 2100 Bioanalyzer and RNA 6000 Nano Kit (Agilent). RNA from three biological replicates was pooled in equimolar amounts and ribosomal RNA was removed by the successive application of the MICROBExpress (Ambion) and RiboZero (Epicentre) kits. The removal of rRNA was subsequently confirmed via the Agilent 2100 Bioanalyzer (RNA 6000 Nano Kit, Agilent). The rRNA-depleted RNA samples were then prepared for sequencing on the Ion Personal Genome Machine (PGM) System as described previously [165]. cDNA libraries were constructed with the Ion Total RNA-Seq Kit v2 (Ion Torrent), and the libraries were used to generate template-positive Ion Sphere Particles (ISPs) with the Ion PGM Template OT2 200 Kit (Ion Torrent). The template positive ISPs were loaded onto Ion 318 v2 chips (Ion Torrent) and sequencing runs were performed with the Ion PGM Sequencing 200 Kit v2 (Ion Torrent). Data analysis was conducted using CLC Genomics Workbench (Qiagen) and the USA300-FPR757 reference genome (accession number: CP000255). RPKM values (Reads Per Kilobase per Million mapped reads) were generated for each gene, a quantile normalization approach applied [166] and a lower limit of 40 RPKM was imposed.

3.3.2.3 Computational metabolic flux prediction

In this study, E-Flux2 (see Chapter 2, section 2.2.3.1) [113] was used to analyze the difference in intracellular metabolic fluxes between the wildtype and the *rpiRc::bursa* mutant. Biomass production was set as the objective function. The transcriptomic data obtained as described in the previous section (section 3.3.2.2) of the two strains were employed for the E-Flux2 analysis. We used *i*SB619 as the genome-scale metabolic model for *S. aureus* USA300 strain [167] with slight modifications in its gene-protein-reaction (GPR) associations. Specifically, since *i*SB619 was constructed based on the closely-related *S. aureus* N315 strain, genes in the model were converted to the corresponding orthologous USA300 genes. The predicted fluxes were normalized by growth rates of the two strains that were measured under the same conditions as the transcriptomic data. The list of metabolic pathways with significant changes is summarized and the full set can be found in Table st4 at the following web address:

http://mbio.asm.org/content/suppl/2016/06/08/mBio.00818-16.DCSupplemental.

3.3.3 Results

In order to screen regulatory genes that may alter leukocidin production, we created a regulator sublibrary from the USA300 Nebraska transposon mutant library collection [168]. This sublibrary consisted of 251 mutants of the JE2 strain, a laboratory version of USA300 LAC, that included gene products with any potential regulatory roles, including ones with nucleotide-binding domains, putative or confirmed HTH motifs, two-component regulatory systems, terminators, and anti-terminators.

All of the leukocidins are known to target human polymorphonuclear leukocytes (hPMNs). Therefore, we screened supernatants collected from the 251 mutants for the ability to lyse hPMNs. Supernatants collected from the sublibrary grown for 3 h were used to intoxicate hPMNs isolated from four human donors. We found several mutants that exhibited altered cytotoxicity for hPMNs. For this study, we chose candidates that showed hypercytotoxicity with the goal of identifying novel repressors involved in the expression of leukocidins.

Compared to wild-type-induced cytotoxicity, mutants with changes in known leukocidin repressors, such as rot, were identified as hypercytotoxic in our screening, validating our assay. Of the mutations that caused increased cytotoxicity, 10 regulators led to neutrophil killing similar to that of a *rot::bursa* mutant (~2.2-fold increased cytotoxicity, Fig. 3.7A). In order to validate these data, we collected supernatants at both 3 and 6 h of bacterial growth. Of the 10 regulators tested, only *rot::bursa* and *rpiRc::bursa* (i.e. the mutation corresponding to NE1142) caused increased cytotoxicity for hPMNs at both time points (Fig. 3.7B and 3.7C). Thus, we decided to focus on characterizing RpiRc and its repressive effects on *S. aureus* virulence.

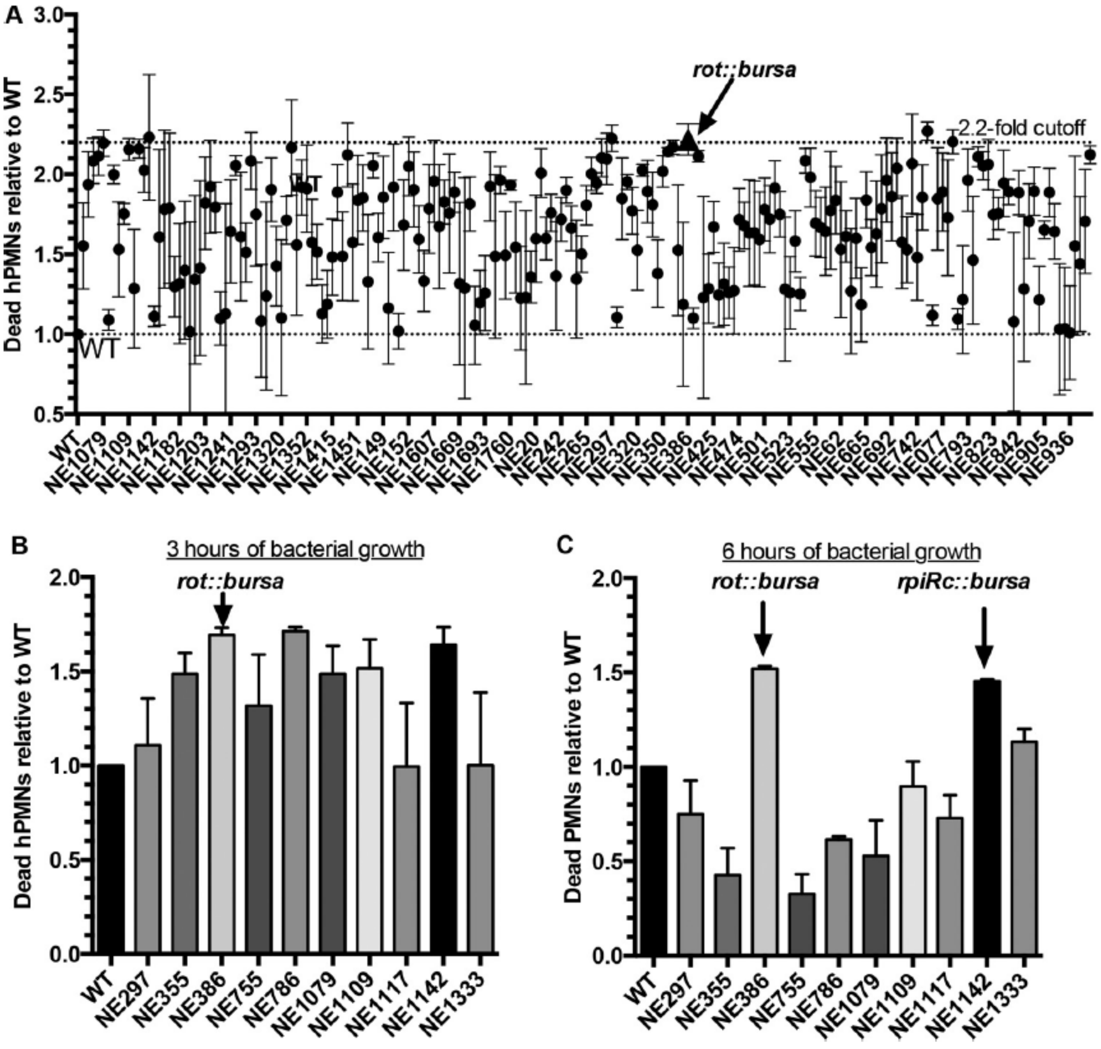


Fig 3.7. Identification of transcriptional regulators that enhance *S. aureus* cytotoxicity. (A) Primary intoxication screening of hPMNs with USA300 JE2 supernatants at a final concentration of 5% (vol/vol). Data points represent neutrophil death caused by an individual mutant relative to that caused by wild-type (WT) bacteria (lower dotted line). Supernatants from each mutant were tested on hPMNs from four donors, and cell viability was measured with CellTiter metabolic dye. A 2.2-fold cutoff was used to identify candidate mutants for further screening (upper dotted line). The data point indicated by the triangle represents the cytotoxicity of a *rot::bursa* mutant. (B and C) Validation intoxication screening of hPMNs isolated from two donors with supernatants from select *S. aureus* mutants. Wild-type and mutant bacteria were grown for 3 (B) and 6 (C) h post-inoculation. Error bars indicate the standard error of the mean.

The RpiR class of proteins has been found in many different bacterial species, including Gram-negative *Escherichia coli* and *Pseudomonas putida* and Gram-positive *Bacillus subtilis* [169–171]. This family of proteins is traditionally thought to include transcriptional regulators involved in sugar metabolism, although their regulons, binding sites, and exact functions are poorly characterized. A recent study identified three RpiR homologs in *S. aureus*, namely, *rpiRa*, *rpiRb*, and *rpiRc* [172].

We first examined the exoproteomes of wild-type and *rpiRc::bursa* mutant strains. The most dramatic difference in abundance was observed in protein bands corresponding to the size of leukocidins (~35 kDa). There were notably higher levels of proteins in that size range in the JE2 *rpiRc::bursa* culture filtrate than in that of wild-type JE2 (Fig. 3.8A), whereas mutations in the other *rpiR* genes had no effect on these toxins.

To demonstrate that the observed phenotype of the JE2 *rpiRc::bursa* mutant was due to the transposon-mediated disruption of *rpiRc*, the mutated allele was transduced into USA300 LAC strain AH1263 (referred to as LAC in this report), another erythromycin-sensitive LAC clone [163]. Compared to the wildtype strain, the isogenic LAC *rpiRc::bursa* mutant also exhibited increased production of proteins that run at the size of leukocidins (Fig. 3.8B). Importantly, this phenotype was fully complemented by the insertion of *rpiRc* in single copy at the SaPI1 attachment site (referred to as the *rpiRc*$^+$ strain in this study) (Fig. 3.8B).

To gain better insight into the effects of RpiRc on exoprotein production in USA300, we analyzed the in vitro culture filtrates by mass spectrometry. Exoproteins were collected from three independent colonies of wild-type LAC and the *rpiRc* mutant strain grown to post-exponential phase, and the protein profiles were analyzed by quantitative mass spectrometry. We observed huge reproducibility among the replicates, as demonstrated by the clustering of the proteins in the heat map (Fig. 3.8C). Among these proteins, we found increased production of most exoenzymes in the *rpiRc* mutant, such as proteases (4- to 6-fold increase), while a few exoenzymes, such as coagulase, were lower in abundance (Fig. 3.8D). Proteins involved in immune evasion and adhesion (including superantigens, Sbi, protein A, and ClfB) were, for the most part, lower in abundance in the *rpiRc* mutant (Fig. 3.8E). We found a stark increase in the production of cytotoxins (leukocidins and phenol-soluble modulins) in the *rpiRc* mutant (Fig. 3.8F). Interestingly, RpiRc seems to differentially regulate the production of leukocidins, as LukSF-PV and LukED were notably upregulated in the *rpiRc* mutant strain, whereas other leukocidins, such as gamma hemolysin and LukAB, were minimally impacted in this strain (Fig. 3.8F).

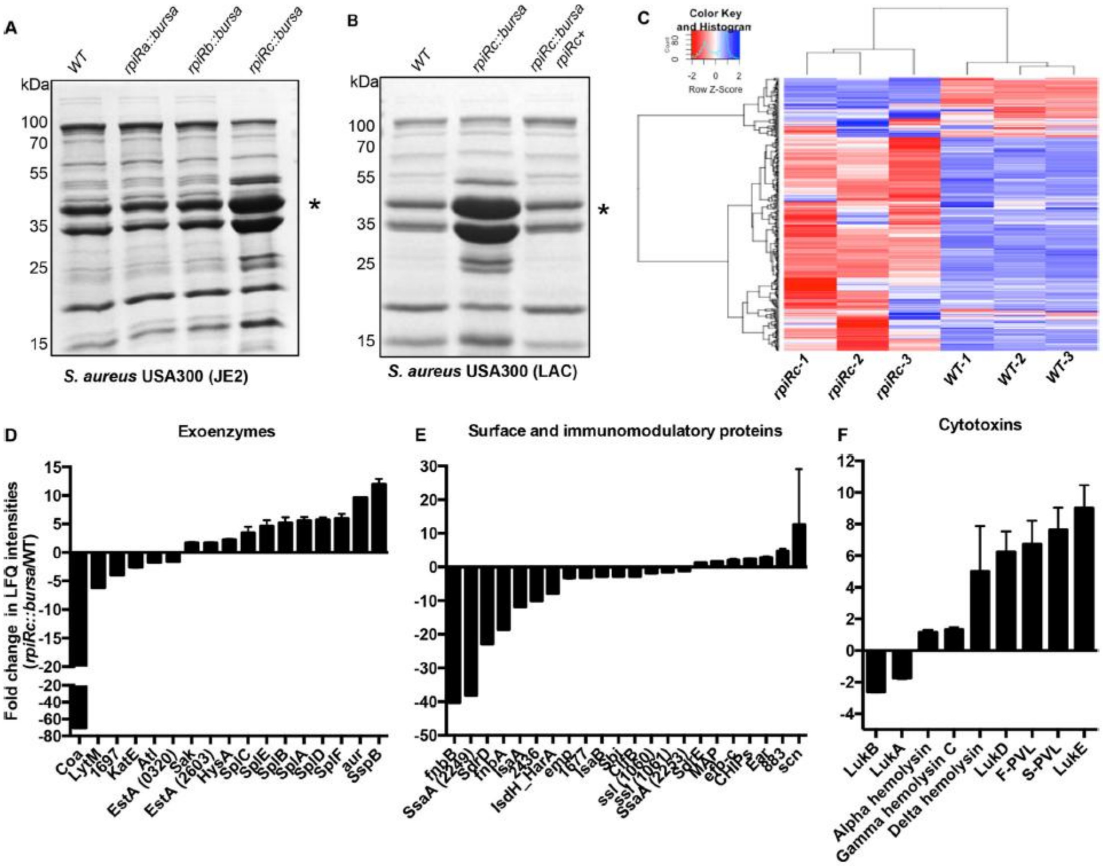


Fig 3.8. RpiRc is a potent regulator of *S. aureus* secreted proteins. (A) Exoprotein profiles of USA300 JE2 wild-type (WT) and *rpiR* mutant bacteria, as assessed by Coomassie staining. The asterisk indicates the approximate leukocidin protein size. (B) Exoprotein profile of USA300 LAC wild-type, *rpiRc*, and *rpiRc*$^+$ isogenic strains. (C) Heat map of LAC wild-type and *rpiRc* mutant secretomes as assessed by mass spectrometry. (D to F) Levels of exoenzymes (D), surface and immunomodulatory proteins (E), and cytotoxins (F) ± the standard deviation in exoproteomes of wild-type versus *rpiRc* mutant USA300 LAC.

The regulon of RpiRc in *S. aureus* is currently unknown. To gain a better understanding of the genes differentially regulated by RpiRc, we performed transcriptome sequencing (RNA-Seq) of RNA isolated from post-exponential phase cultures of wild-type and *rpiRc* mutant USA300. Genes that were 5-fold up- or down-regulated in the *rpiRc::bursa* mutant are shown in Fig. 3.9A. Of these, the lukED and lukSF-PV leukocidin transcripts were drastically upregulated. In addition, genes involved in capsular polysaccharide biosynthesis and some proteases, such as *sspABC*, were also upregulated. In contrast, many genes involved in sugar metabolism and transport and surface virulence proteins (such as spa) were downregulated.

Since RpiRc is a regulator of metabolic enzymes, we wanted to determine if there were any general trends in the up- or downregulation of certain metabolic pathways. To address this question, we performed *in silico* metabolic profiling analyses of the transcriptomic data [46,47,113]. As shown in Fig. 3.9B (see Table st4 at http://mbio.asm.org/content/suppl/2016/06/08/mBio.00818-16.DCSupplemental), we observed that three clusters of metabolic pathways were differentially activated in wild-type and *rpiRc* mutant bacteria. First, and notably, we observed a signature of decreased tricarboxylic acid (TCA) cycle activity in the *rpiRc* mutant. Second, and in contrast, several amino acid metabolic pathways were more active in the mutant than in the wild type. Third, we observed a slight but consistent activation of the glycolysis and gluconeogenesis pathways. While the fold differences in metabolic fluxes in glycolysis/gluconeogenesis between wild-type and *rpiRc* mutant bacteria are only moderate, we nevertheless observed that many genes in these two pathways were upregulated in the *rpiRc* mutant (Fig. 3.9B). Of note, we observed no growth defect in vitro when comparing the wild-type strain and the isogenic strain lacking *rpiRc*, consistent with observations reported previously [172]. Taken together, mutation of *rpiRc* seems to lead to decreased activity of the TCA cycle and an increase in glycolysis/gluconeogenesis and certain amino acid biosynthetic pathways. In UAMS-1, RpiRc was observed to increase the expression and activity of some PPP genes [172]. In our analyses, while there were no notable PPP shifts, we observed potential positive regulatory roles of RpiRc in the TCA cycle.

In consistent with the RNA-Seq analyses, proteomic analyses showed a striking increase in the expression of lukSF-PV and lukED in the mutant (Fig. 3.9C), further validating that RpiRc is involved in the expression of these specific leukocidin-coding genes.

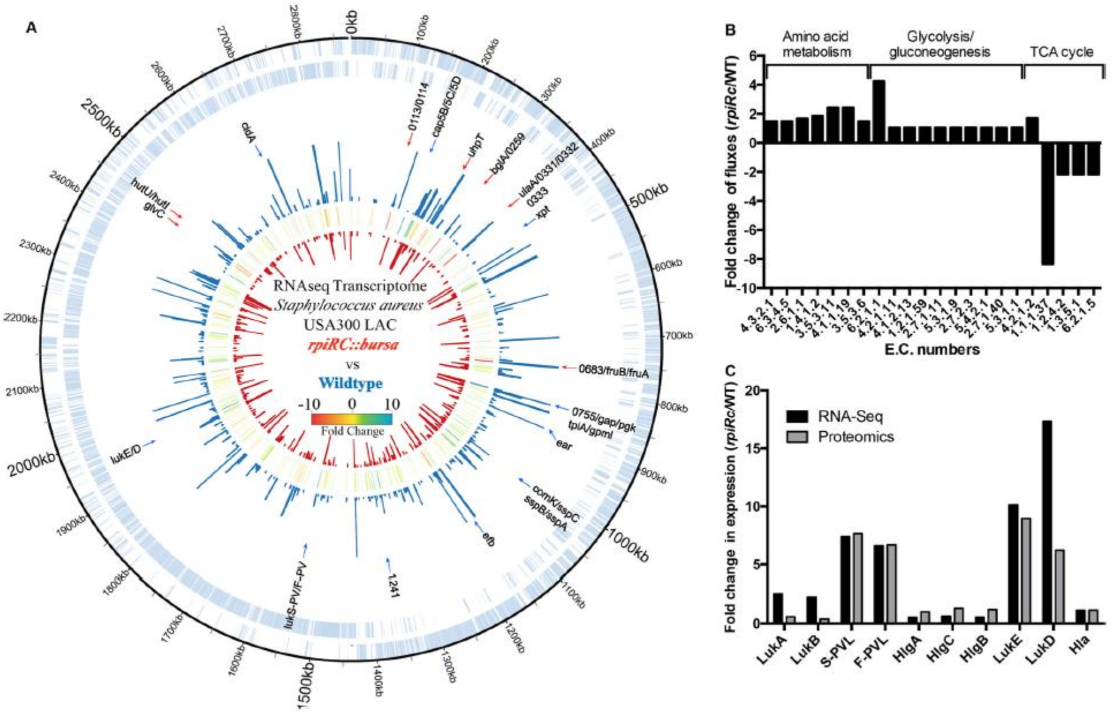


Fig. 3.9. Defining the RpiRc regulon and the RpiRc-associated metabolic changes. RNA-Seq of USA300 wild-type and *rpiRc* mutant bacteria. (A) Genomic map depicting transcription profiles. The outer circle (blue) indicates the RPKM (number of reads per kilobase per million mapped reads) of the wild type, and the inner circle (red) indicates the RPKM of the *rpiRc* mutant. The center circle is a heat map showing the fold differences in expression. Genes that were 5-fold upregulated (blue arrows) or 5-fold downregulated (red arrows) are indicated. (B) Fold differences in metabolic flux between the wild type (WT) and the *rpiRc* mutant based on RNA-Seq results. The enzyme classification (E.C.) numbers represent specific enzymatic reactions and their corresponding pathways. (C) Fold differences in transcript (RNA-Seq) and protein (proteomics) abundance of the individual leukocidins between wild-type and *rpiRc* mutant bacteria.

3.3.4 Discussion

S. aureus secretes a diverse array of virulence factors (47), including the bicomponent pore-forming leukocidins (3) that target leukocytes. Studying the leukocidin mode of action, as well as the regulatory networks that govern their expression, will contribute to the understanding of *S. aureus* pathogenesis. Here, we undertook a screening strategy to identify potential regulators involved in the control of leukocidin expression (Fig. 3.7). We identified RpiRc, a metabolic transcriptional regulator, as a potent repressor of leukocidins (Fig. 3.8). The wild-type and *rpiRc* mutant strain exoprotein and transcriptomic profiles described here provide additional insight into the function of RpiRc in *S. aureus*. In the RNA-Seq analyses, in addition to the leukocidins, we observed >2-fold changes in genes involved in metabolism, sugar transport, translation, transcription, TCSs, and other regulators. In the computational metabolic flux analysis, the activity of the TCA cycle, glycolysis/gluconeogenesis, and certain amino acid biosynthetic pathways seems to be affected by the mutation of *rpiRc* (Fig 3.9). While we did not validate all of the candidate genes or pathways, it is likely that RpiRc has direct or indirect effects on sugar metabolism. Validation and characterization of these targets may indicate other roles for RpiRc in *S. aureus* and may provide clues to links between metabolism and virulence.

To be a successful pathogen, *S. aureus* has to adapt to the harsh environments encountered within the host. In recent years, interest in *S. aureus* metabolism has reemerged, as distinct links between metabolism and pathogenesis are increasingly identified [173,174]. In *S. aureus*, several transcriptional factors sense metabolites and regulate virulence in response to these signals. Examples of metabolite-sensing regulators in *S. aureus* are CcpA and CcpE. These carbon catabolite repressors sense glycolytic intermediates, and in addition to regulating uptake of nutrients such as glucose, they also regulate the synthesis of virulence factors [175,176]. Another well-studied nutrient sensor is the CodY transcriptional regulator, which responds to branched-chain amino acids and GTP in *S. aureus* [177]. In response to nutrient availability, CodY regulates the synthesis of alpha-toxin and certain adhesins via the *agr* system [178]. Importantly, inactivation of many of these regulators (including the ones cited above) alters *S. aureus* pathogenesis, supporting the notion that metabolism is intimately linked with the pathogenic lifestyle of this bacterium. The data presented here support the idea that RpiRc is a critical transcriptional regulator that may respond to various environmental conditions to increase, decrease, or fine-tune *S. aureus* virulence.

CHAPTER 4: Assessment of our methods in cells grown on different substrates

4. 1 Background

As shown in Chapter 3, computational tools that predict system-level and condition-specific metabolic flux distributions by integrating transcriptomic data in genome-scale metabolic models have many useful applications. For this reason, many methods for inferring metabolic fluxes from gene expression data have been, and continue to be, developed [32,57]. However, absolute and comparative performance of these methods remain poorly understood partly due to the lack of experimental data for validation (i.e. transcriptome measurements coupled with corresponding intracellular flux measurements).

Including our work described in Chapter 2, there are previous studies [32,113] that extensively evaluated the predictive capability of these methods using measured extracellular and intracellular fluxes in multiple experimental conditions. In these studies, however, the validation was performed exclusively on the datasets of *E. coli* and *S. cerevisiae* cells grown on glucose as the sole carbon source. Not surprisingly, given that the cells were cultured on the same substrate and that metabolic pathways responsible for metabolism of glucose are conserved among different species [179], there are significant similarities in the metabolic flux distribution across all validation datasets.

Which metabolic pathways get employed by a cell are influenced by what kind of carbon sources are being taken up by the cell. For example, in the natural environment, heterotrophic microorganisms are exposed to a wide set of possible carbon sources that can support their growth such as sugars, polyols, alcohols, organic acids, and amino acids

[180]. The carbon substrates whose metabolism have been most widely studied include hexose monosaccharides (e.g. glucose, fructose, mannose or galactose), disaccharides (e.g. sucrose or maltose), and compounds with two carbons (e.g. ethanol or acetate). The metabolic reactions activated for metabolizing hexoses and disaccharides share similar pathways, differing in the initial steps of metabolism, and most metabolic building blocks are derived from intermediate metabolic enzymes of glycolysis, the TCA cycle, and the pentose phosphate pathway. Significant changes are observed, however, in the metabolism of two-carbon compounds compared to that of sugars since gluconeogenesis and the glyoxylate cycle are essential for converting two-carbon compounds into anabolic precursors [180]. The metabolism of animal cells is more complex partly because they require multiple nutrients at the same time, which is one of the reasons why culturing them is more difficult than microorganisms. In addition to a main carbon source such as glucose, animal cells grown in culture must be supplied with nine essential amino acids (i.e. histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine) that cannot be synthesized by adult vertebrate animals [181]. In the case of autotrophs, they gain energy and chemical building blocks from inorganic carbon courses (e.g. CO_2) by employing carbon fixation pathways. The Calvin cycle is one of the well-known carbon fixation pathways found in many photoautotrophic organisms [182].

Given the impact of different carbon sources on the usage of metabolic pathways in various organisms, it is desirable to assess performance of E-Flux2 and SPOT and other computational methods of a similar nature using a new dataset obtained from cells grown on various carbon sources so as to evaluate predictive power and demonstrate generality of these methods. To this end, we have complied additional 23 experimental conditions of transcriptome measurements coupled with corresponding central carbon metabolic intracellular flux measurements (8 in *E. coli*, 8 in *Bacillus subtilis*, 3 in *Synechocystis* sp. PCC 6803, 2 in *Synechococcus* sp. PCC 7002, and 2 in H4IIE rat hepatoma cell line). Using this new dataset, the performance of our methods and a representative sample of competing methods has been further validated.

4. 2 Materials and methods

4.2.1 Gene expression and flux dataset, and metabolic models used for this study

4.2.1.1 Data and model for *E. coli*

For *E. coli*, both the measured gene expression and the flux data were obtained from a previous study by Gerosa et al.[183]. In this study, data were measured from *E. coli* wild-type BW25113 cells growing exponentially on eight different carbon sources: glucose, galactose, gluconate, fructose, glycerol, pyruvate, acetate, and succinate. We used *i*JO1366 [184] as the genome-scale metabolic model.

4.2.1.2 Data and model for *B. subtilis*

For *B. subtilis*, we used transcriptomic and flux data published in [185] and [186], respectively. Data were obtained from *B. subtilis* BSB168 cells grown under eight conditions defined by different carbon sources: glucose, fructose, gluconate, succinate + glutamate, glycerol, malate, malate + glucose, and pyruvate. For the genome-scale metabolic model of *B. subtilis*, the model published by Oh et al. [187] was used.

4.2.1.3 Data and model for *Synechocystis* sp. PCC 6803

For *Synechocystis* sp. PCC 6803, transcriptomic data was graciously provided by Dr. Le You (University of California San Diego, USA) and Dr. Yinjie Tang (Washington University in St. Louis, USA) [188]. The flux data was compiled from three different publications [188–190]. Data were measured from the strain *Synechocystis* sp. PCC 6803 grown under three different conditions: photoautotrophic (i.e. CO_2 as the main carbon source) [189], photomixotrophic (i.e. CO_2 + glucose) [190], and heterotrophic (i.e. glucose) [188], respectively. We used the genome-scale metabolic model of *Synechocystis* sp. PCC 6803 developed by Knoop et al. [130].

4.2.1.4 Data and model for *Synechococcus* sp. PCC 7002

For *Synechococcus* sp. PCC 7002, the transcriptomic data was obtained from a previous publication by Ludwig & Bryant [142]. The flux data was kindly provided from Dr. Xiao Qian (Rutgers University, USA). Data were measured from *Synechococcus* sp. PCC 7002 cells grown photoautotrophically with and without nitrate. *i*Syp821[93] was used for the organism's genome scale-metabolic model.

4.2.1.5 Data and model for H4IIE rat hepatoma cell

For H4IIE rat hepatoma cell, the gene expression data was collected from a previous study by Sriram et al. [191]. The corresponding flux data was obtained from another publication of the same research group [192]. Data were measured from wild-type and glycerol kinase (GK) – overexpressing H4IIE rat hepatoma cell lines. In addition to the main carbon source, glucose, amino acids were supplied to these mammalian cell cultures. Due to the unavailability of a rat-hepatoma cell line-specific genome-scale metabolic model, we converted a human hepatocyte model, HepatoNet1 [193] to a rat-hepatocyte one according to the way described in [194]

4.2.2 Computational metabolic flux prediction

In this study, E-Flux2 (see Chapter 2, section 2.2.3.1), SPOT (section 2.2.3.2) [113], Lee's method [48], and pFBA [81] were used to predict metabolic flux distributions. Biomass production was set as the objective function for E-Flux2 and pFBA. All methods used in this study implemented in MATLAB (The Mathworks, Inc., Natick, Mass.) are provided with their original publications [48,81,113]. Analyses were performed using MATLAB R2013b with Gurobi Optimizer 5.6 (Gurobi Optimization, Inc., Houston, Texas). SBMLToolbox was used to convert an SBML (Systems Biology Markup Language) model into a MATLAB data structure [84]. Computations were carried out on the Window 8 platform using a personal computer with an Intel Core i5 3.10 GHz processor with 8 GB of RAM. E-Flux2 and SPOT methods are also implemented in MOST (Metabolic Optimization and Simulation Tool) which is available at http://most.ccib.rutgers.edu [85].

4.2.3 Validation of the predictive accuracy of the methods used in this study

The predictive accuracy of all methods were validated by calculating the uncentered Pearson product-moment correlation between *in silico* fluxes and corresponding ^{13}C-determined *in vivo* intracellular fluxes as previously described in [113], that is

$$\frac{v_p \cdot v_m}{\|v_p\|\|v_m\|}$$

where v_p and v_m are the predicted and measured vectors of intracellular fluxes, respectively, and $\|\cdot\|$ denotes the l^2 norm. The uncentered Pearson correlation is a scale-independent metric of linear relationship, hence, is a good metric of the performance of flux inference methods because fluxes predicted by these methods have an unknown scale factor. A value of the correlation coefficient close to +1 or -1 indicates a strong positive or negative linear relationship between v_p and v_m, respectively. A value of 0 indicates no linear relationship [82].

If a measured reaction corresponds to the set of consecutive reactions in the model that are linked with intermediate metabolites (AND relationship), then the minimum flux value—the slowest reaction rate—among those predicted fluxes was used to calculate correlation with the corresponding measured flux since the rate of a reaction with several steps is determined by the slowest step, which is known as the rate-limiting step in chemical kinetics [83]. If a measured flux corresponds to multiple identical reactions (OR relationship), the sum of those predicted fluxes was used to calculate the correlation since the rate of a reaction would be faster, that is, would have greater flux value, as the number of reactions that can perform an identical chemical conversion increases.

4. 3 Results

In *E. coli* and yeast cells grown on glucose, E-Flux2 and SPOT performed well in predicting intracellular central carbon metabolic flux distributions over a representative sample of competing methods [113]. To test generality of E-Flux2 and SPOT, we evaluated predictive accuracy of them by calculating the uncentered Pearson correlation between predicted fluxes and measured intracellular fluxes (see section 2.2.4 or section 4.2.3 for details), on the newly compiled 23 experimental conditions of transcriptomic data coupled with corresponding measured flux data. The dataset consists of 8, 8, 3, 2, and 2 conditions of *E. coli*, *B. subtilis*, *Synechocystis* sp. PCC 6803, *Synechococcus* sp. PCC 7002, and H4IIE rat hepatoma cell line data, respectively (refer to section 4.2.1 for carbon source information). The average of 23 correlations that were produced by each of our methods was compared with those of two competing methods, Lee's method [48] and pFBA [81] considered as leading algorithms for predicting metabolic flux distributions with and without integrating transcriptomic data, respectively [32,48,57,81,113].

The results are shown in Fig 4.1. Looking at the average of the correlation values, we see that pFBA (average correlation: 0.64) performs slightly better than the other three methods. In consideration of the spread of the values, however, SPOT seems to provide more stable predictions across different experimental conditions (Fig 4.1).

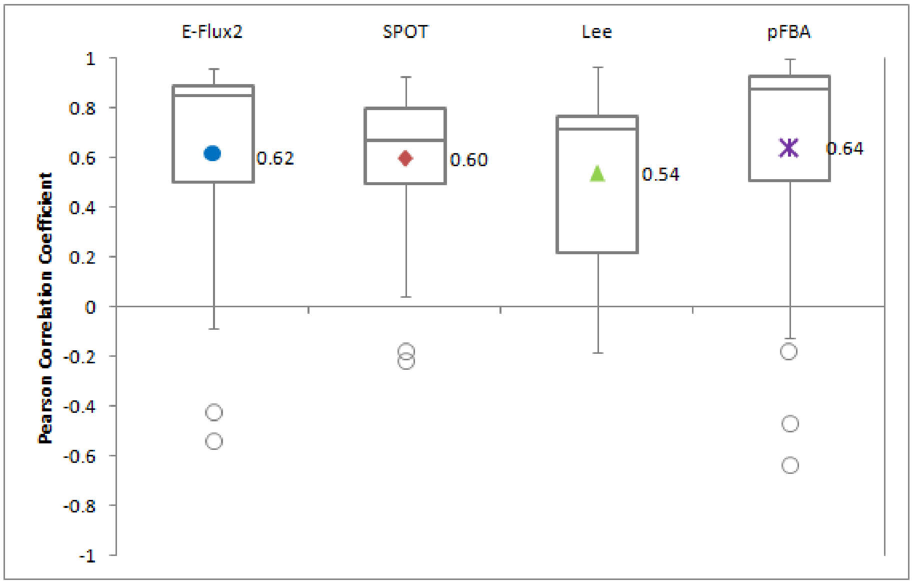


Fig 4.1 The average and the distribution of the Pearson correlations between measured and predicted fluxes produced by each of the four methods for the 23 experimental conditions of dataset newly compiled. Boxes represent the interquartile range (IQR) between first and third quartiles. The center line inside each box shows the median. The blue, red, green, and purple markers represent the averages of E-Flux2, SPOT, Lee's method, and pFBA, respectively. Whiskers extend 1.5 × IQR from the 25th and 75th percentiles, respectively. Outliers beyond the whiskers are represented by grey open circles. Sample size n = 23 for each method.

To identify at which conditions each method performs well or not, we examined and compared detailed correlation values of the four methods for all 23 conditions (see Fig 4.2). We see that all methods yield good or moderate predictions in *E. coli* and *B. subtilis* with an average correlation around 0.8, regardless of which carbon source was available during growth. In particular, in agreement with our prior findings for *E. coli* and *S. cerevisiae* grown on glucose [113], E-Flux2 (blue circle in Fig 4.2) and SPOT (red diamond) work well in *E. coli* and *B. subtilis* cultured on glucose (see conditions 5 and 10 in Fig 4.2). However, the average correlation of all methods drops, and variance increases, for the two cynobacterial strains (i.e. *Synechocystis* sp. PCC 6803 and *Synechococcus* sp. PCC7002) and for the rat hepatoma cell line, H4IIE cells.

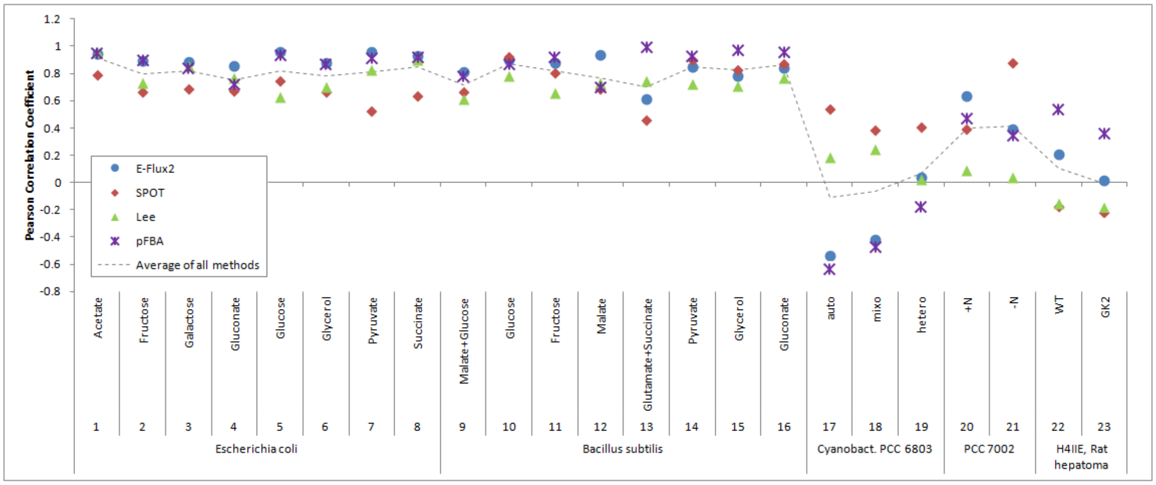


Fig 4.2 Detailed correlations between measured and predicted fluxes of the four methods for 23 experimental conditions. The blue circle, red diamond, green triangle, and purple asterisk markers indicate the Pearson correlation between measured and predicted fluxes of E-Flux2, SPOT, Lee's method, and pFBA, respectively. The dash line shows the average correlation of all methods per each condition.

After observing that predictive power of each method is influenced less by specific carbon source but more by organism, we revisited previous results on the 20 experimental conditions of *E. coli* and yeast dataset compiled for the study in Chapter 2. In this dataset, E-Flux2 and SPOT outperformed the other methods in predicting intracellular metabolic flux distributions especially of yeast cells (conditions 35 to 43 in Fig 4.3).

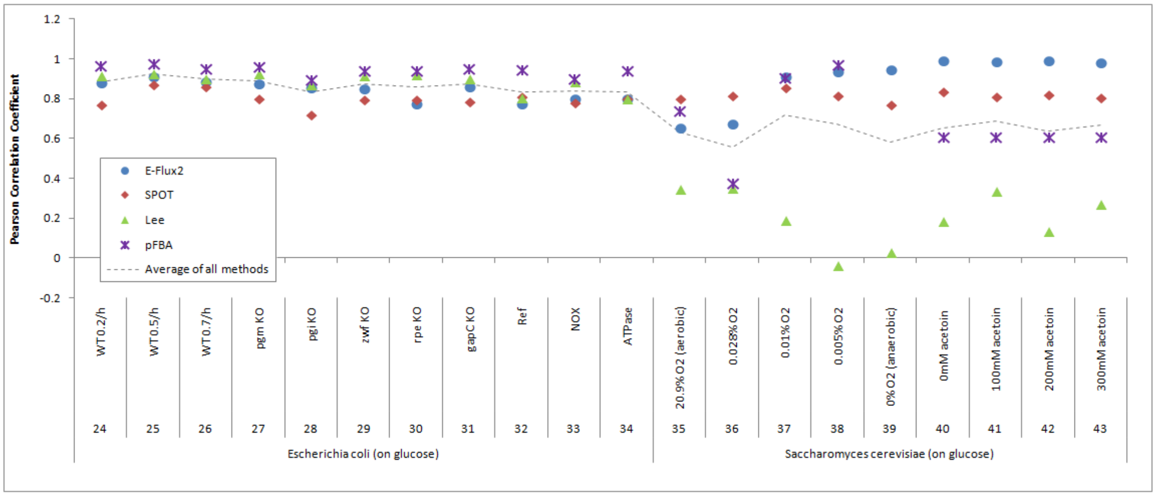


Fig 4.3 Detailed correlations between measured and predicted fluxes of the four methods for the 20 experimental conditions of dataset previously compiled for the study in Chapter 2. The data were obtained from *E. coli* and yeast cells grown on glucose as the main carbon source (see Chapter 2, section 2.2.1 for more information). The blue circle, red diamond, green triangle, and purple asterisk markers indicate the Pearson correlation between measured and predicted fluxes of E-Flux2, SPOT, Lee's method, and pFBA, respectively. The dash line shows the average correlation of all methods per each condition.

In general, the larger the dataset is used for validation, the more reliable the conclusions that can be derived. For this reason, we combined the 20 experimental conditions of the dataset previously complied [113] and the 23 conditions of the one newly compiled. The distribution and the average of the Pearson correlations between measured and predicted fluxes generated by each method for all 43 conditions is shown in Fig 4.4.

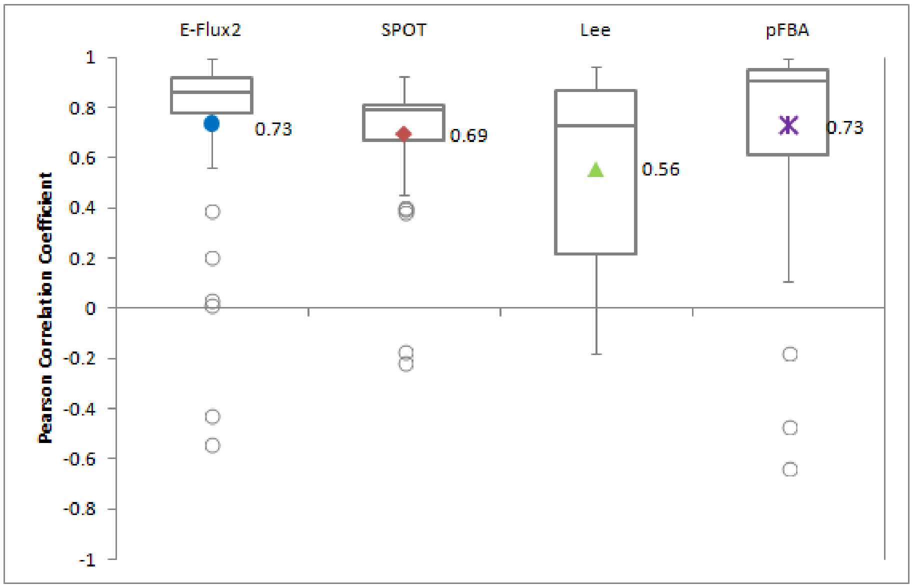


Fig 4.4 The average and the distribution of the Pearson correlations between measured and predicted fluxes produced by each of the four methods for the 43 experimental conditions of dataset (i.e. 23 conditions of dataset newly compiled + 20 conditions of dataset previously compiled). Boxes represent the interquartile range (IQR) between first and third quartiles. The center line inside each box shows the median. The blue, red, green, and purple markers represent the averages of E-Flux2, SPOT, Lee's method, and pFBA, respectively. Whiskers extend 1.5 × IQR from the 25th and 75th percentiles, respectively. Outliers beyond the whiskers are represented by grey open circles. Sample size n = 43 for each method.

As shown in Fig 4.4, with regards to the average correlation, it seems that E-Flux2 and pFBA (average correlation of both methods: 0.73) perform best, followed by SPOT (average correlation: 0.69), then Lee's method (average correlation: 0.56). In light of the mid-spread (i.e. the range of the correlations of the middle 50% of them, which corresponds to the height of each box in Fig 4.4), however, E-Flux2 and SPOT seem to provide more stable predictions than pFBA across 43 different experimental conditions.

Unicellular heterotrophic microorganisms such as *E. coli* and yeast offer attractive systems for biological and industrial applications because of their fast growth rate and their relative simplicity of genetic manipulation. If our methods perform well in those single-celled heterotrophic microbes, they will be already useful in a wide range of areas even if they do not perform best in every organism. On account of this, we drew a box plot (Fig 4.5) and a dot plot (Fig 4.6) using only the 36 (out of 43) experimental conditions of data obtained from unicellular heterotrophic microorganisms (i.e. *E. coli, S. cerevisiae*, and *B. subtilis*) to examine predictive performance of our methods in those organisms.

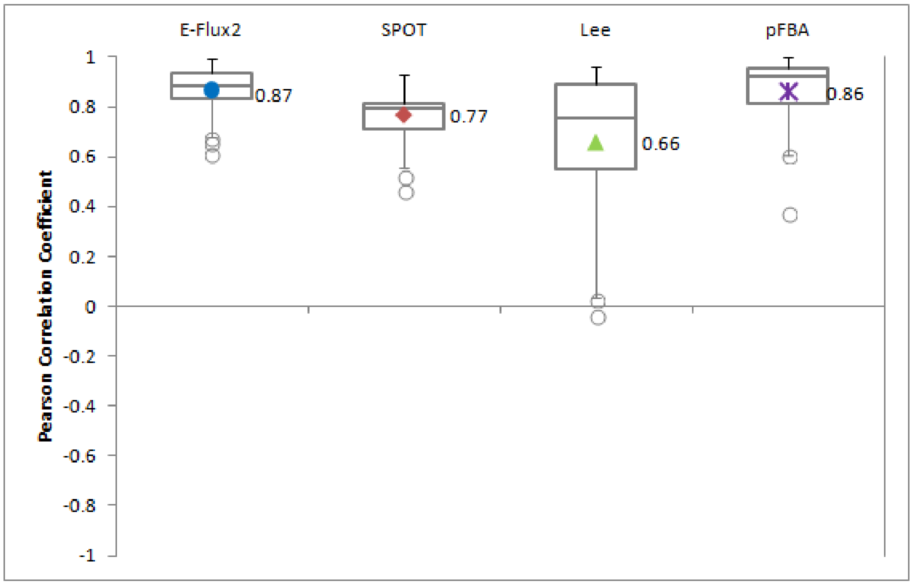


Fig 4.5 The average and the distribution of the Pearson correlations between measured and predicted fluxes produced by each of the four methods for the 36 experimental conditions of dataset of unicellular heterotrophic microorganisms. Boxes represent the interquartile range (IQR) between first and third quartiles. The center line inside each box shows the median. The blue, red, green, and purple markers represent the averages of E-Flux2, SPOT, Lee's method, and pFBA, respectively. Whiskers extend 1.5 × IQR from the 25th and 75th percentiles, respectively. Outliers beyond the whiskers are represented by grey open circles. Sample size n = 36 for each method.

As shown in Fig 4.5, both E-Flux2 and pFBA predicted intracellular central carbon metabolic flux distributions well with an average correlation of 0.87 and 0.86, respectively. Even though pFBA has a slightly lower average correlation than E-Flux2, it has the highest median. However, considering pFBA's outliers in yeast (Fig 4.6), the highest 75th percentile and the highest average of E-Flux2, E-Flux2 appears to have smaller variability than pFBA. Overall, it seems that there is no method which is significantly superior to the other methods in all cases, which will be further discussed in the next section.

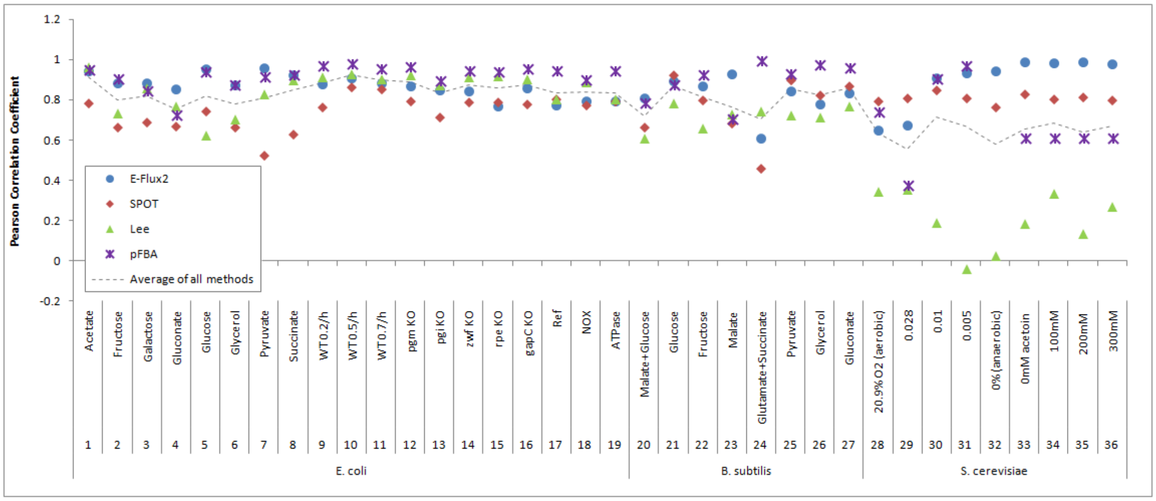


Fig 4.6 Detailed correlations between measured and predicted fluxes of the four methods for the 36 experimental conditions of dataset of unicellular heterotrophic microorganisms. The data were obtained from *E. coli*, *S. cerevisiae*, and *B. subtilis* cells. The blue circle, red diamond, green triangle, and purple asterisk markers indicate the Pearson correlation between measured and predicted fluxes of E-Flux2, SPOT, Lee's method, and pFBA, respectively. The dash line shows the average correlation of all methods per each condition.

4. 4 Discussion

In this study, we have additionally compiled 23 experimental conditions of transcriptome measurements along with corresponding flux measurements to evaluate predictive power of E-Flux2 and SPOT in cells grown on various nutrients. As shown with the dash line in Fig 4.2, it seems that predictions of each method are influenced less by specific carbon source but more by organism. Then we added more data (i.e. the 20 experimental conditions of dataset previously compiled for the study in Chapter 2) to the study to get more reliable insights on predictive accuracy of our methods. Importantly, the total 43 experimental conditions of dataset we have assembled is the largest to date for validating methods for predicting metabolic flux distributions from transcriptomic data.

As can be seen from the dash lines in Figs 4.2 and 4.4, all methods we have tested (i.e. E-Flux2, SPOT, Lee's method, and pFBA) produce good predictions with an average correlation around or above 0.8 in *E. coli* and *B. subtilis*. However, overall predictions of the methods are poorer, and the variance in predictive performance among these methods is larger in yeast, the two cynobacterial strains, and the rat hepatoma cell line. This organism-specific variation in predictions may be attributed to the fact that the former two organisms (i.e. *E. coli* and *B. subtilis*) are relatively simple and well-studied microorganisms compared to the latter group of organisms. It is likely, therefore, that the genome-scale metabolic models of *E. coli* and *B. subtilis* are more complete thanks to more accurate genome annotations, more detailed and correct gene-protein-reaction (GPR) associations, and abundant biochemical literature essential for reconciling the model.

In the case of the organisms where there is large variation in the predictive performance of the four methods (i.e. *S. cerevisiae*, *Synechocystis* sp. PCC 6803 and *Synechococcus* sp.

PCC7002, and H4IIE rat hepatoma cell line), we were able to observe that integration of transcriptomic data sometimes improves metabolic flux predictions and sometimes not. For example, first, the outperformance of E-Flux2 over pFBA in yeast (conditions 35 to 43 in Fig 4.3) suggests that maximization of biomass production (i.e. the objective function used for E-Flux2 and pFBA) is a proper objective function for studying yeast metabolism, and integrating gene expression information can further increase accuracy in predictions. Second, SPOT seems to provide better predictions than the other methods in the two cyanobacterial strains (conditions 17 to 21 in Fig 4.2). This implies that the maximization of the correlation between gene expression and flux distributions might be a better objective function than the conventional one of maximizing biomass yield for the study of cyanobacterial metabolism. The biological reason behind this observation might be due to a dichotomy between autotrophs and heterotrophs regarding the strength of homeostatic regulation [195]. Autotrophic organisms are considered to be relatively plastic because their stoichiometric composition can vary widely with fluctuations in nutrient supply and light, while heterotrophs are generally thought to be strictly homeostatic [196]. It is possible that metabolic flux changes in autotrophic organisms are more susceptible to the abundance of transcripts continuously being adjusted in response to ever-changing environments. Third, on the other hand, pFBA seems to predict better than the other methods in the H4IIE rat hepatoma cell line. It seems that the objective function of maximizing biomass production can also be applied to study metabolism of more complex organisms such as mammalian cancer cells. The poor performance of the other three methods might be due to incorrect GPR associations in the rat hepatocyte model which was converted from a human model; Or, it is possible that inferring metabolic fluxes from gene expression data is more challenging in higher organisms because of their more complicated regulatory mechanisms. What should be noticed here is that, the results of the cyanobacterial and rat hepatoma cells should be carefully interpreted since only 5 and 2 experimental conditions of data were tested, respectively. We need more data to draw firm conclusions regarding predictive power of the methods tested in cyanobacteria and rat hepatoma cell line. Still, to the best of our knowledge, this is the first study that evaluated predictive accuracy of methods for inferring metabolic fluxes from transcriptomic data in autotrophs and mammalian cells.

In a previous study by Machado and Herrgård [32], pFBA outperformed various methods that infer metabolic fluxes from transcriptomic data, which has cast doubt on the necessity of utilizing gene expression data in constraint-based modeling. pFBA produced good predictions in the dataset we tested as well. However, pFBA was not remarkably better than our methods. Our methods, i.e. E-Flux2 and SPOT, yielded essentially as good predictions as pFBA in terms of the average correlation (see the result of E-Flux2 in Fig 4.5) or overall stability in predictions (see the result of SPOT in Fig 4.4) -- unlike a representative sample of previous methods of a similar nature that were used for comparison in the previous report (e.g. E-Flux, Lee's method). The main difference between pFBA and our methods is the input information required. Our methods need transcriptome measurements, and pFBA needs knowledge of an objective function and the specification several fluxes such as nutrient(s) (e.g. glucose and oxygen as limiting organic and inorganic nutrients, respectively) uptake rate(s) [53]. If information required for pFBA are available and getting it is less demanding than measuring transcriptomic data, pFBA seems to be a good choice for predicting metabolic flux distributions. For instance, it might be simpler to use pFBA to estimate metabolic flux distributions of cells grown on a single carbon source in minimal medium. However, those information that are needed to run pFBA are not always available. It may be easier to measure transcriptomic data than to quantify extracellular or intracellular fluxes especially when cells are grown on multiple nutrients, in an undefined medium, or under natural conditions. Therefore, the choice of method to use is a matter of trade-off between the amount of information required for simulation and some absolute flux information. To sum up, the results shown in this study demonstrate that our methods can be a good alternative to pFBA especially when studying metabolism of unicellular heterotrophic microorganisms or when the information needed as input for pFBA is unavailable.

CHAPTER 5: Conclusion

As stated in Chapter 1 section 1.5, the first of the three objectives of this study was to develop new computational methods, for inferring system-level and condition-specific metabolic flux distributions from transcriptomic data, that overcome shortcomings of existing methods. We have achieved this objective, as demonstrated in Chapter 2, by developing two computational methods called E-Flux2 and SPOT, to be employed when a suitable biological objective is available and unavailable, respectively. Our method combines all desirable characteristics summarized in Table 2.1 including applicability to a wide range of experimental conditions and production of a unique solution. Its features also include fast running time and the availability of a user-friendly implementation at http://most.ccib.rutgers.edu/.

The second objective study was to apply the newly developed methods to solve biological research problems, demonstrating their usefulness in biology. We have accomplished the second objective, which is covered in Chapter 3. We have shown that our methods can be used to understand the lipid accumulation mechanism of nitrogen-stressed *Phaeodactylum tricornutum* cells and provide a metabolic engineering strategy to further enhance their lipid biosynthesis (section 3.1), verify the predictive power of a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002 (section 3.2), and examine the metabolic impacts of RpiRc, a potent repressor of microbial toxins in *S. aureus* (section 3.3).

The last objective of this study was to test the generality of the usefulness of our methods through extensive validation with massive experimental data, and making the dataset used for validation publicly available for the research community in this field. As shown in Chapter 2 and Chapter 4, we have validated the predictive accuracy of our methods and compared it with that of other competing methods (e.g. Lee's method and pFBA, which are considered as leading algorithms for predicting metabolic flux distributions with and without integrating transcriptomic data, respectively) using a total of 43 experimental conditions of transcriptome measurements coupled with corresponding central carbon metabolic intracellular flux measurements (19 in *E. coli*, 9 in *S. cerevisiae*, 8 in *B. subtilis*, 3 in *Synechocystis* sp. PCC 6803, 2 in *Synechococcus* sp. PCC 7002, and 2 in H4IIE rat hepatoma cell line). This is the largest and most comprehensive dataset compiled to date, which we hope will contribute to the resources for future development of methods of a similar nature. To the best of our knowledge, even though the number of data points tested is small, this is the first study that evaluated predictive accuracy of methods for inferring metabolic fluxes from transcriptomic data in autotrophs and mammalian cells.

We have shown that predictions of each method are influenced less by specific carbon source but more by organism (Chapter 4, Fig 4.2). This organism-specific differences may be attributed to the level of completeness of the genome-scale metabolic model, complexity of regulatory mechanisms that influence susceptibility of metabolic flux distributions to transcriptional changes, or difficulty in accurately measuring experimental data.

In a previous study by Machado and Herrgård [32], pFBA outperformed various methods that infer metabolic fluxes from transcriptomic data, which has cast doubt on the utility of gene expression data in constraint-based modeling. In the 43 experimental conditions we tested, however, there was no method which is significantly superior to the other methods in all cases (Chapter 4, Fig 4.4). Our methods, i.e. E-Flux2 and SPOT, were able to provide essentially as good as or better predictions than pFBA in terms of the average correlation especially in unicellular heterotrophic microorganisms (Chapter 4, the result of E-Flux2 in Fig 4.5) or overall stability in predictions (Chapters 2 and 4, the result of SPOT in Figs 2.5 and 4.4). The main difference between pFBA and our methods is the input information required for simulation; pFBA requires knowledge on a proper objective function and specifying several fluxes such as nutrient(s) uptake rate(s). If information required for pFBA are available and getting it is less demanding than measuring transcriptomic data, pFBA seems to be a good choice for predicting metabolic flux distributions. On the other hand, it may be easier to measure transcriptomic data than to quantify fluxes especially when cells are grown on multiple nutrients, in an undefined medium, or under natural conditions. Hence, the choice of method to use is a matter of trade-off between the amount of information required for simulation and some absolute flux information.

This research has achieved the three main objectives outlined in the beginning of this dissertation, and based on the findings, we can draw a conclusion that the methods that we developed for inferring metabolic flux distributions from transcriptomic data can be useful especially in studying metabolism of single-celled heterotrophic organisms and/or in the absence of information required as input for standard FBA or pFBA.

The summary of this research is depicted by Fig 5.1 below.

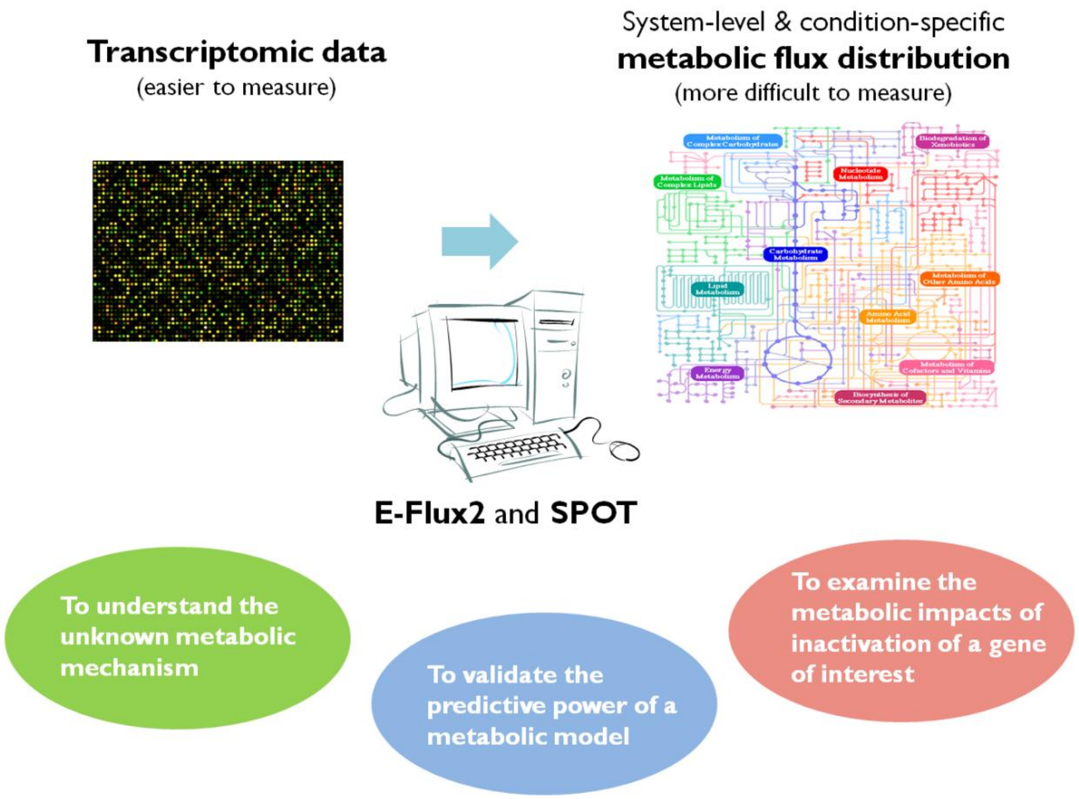

Transcriptomic data
(easier to measure)
System-level & condition-specific
metabolic flux distribution
(more difficult to measure)
Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Metabolism of Complex Lipids
Nucleotide Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Metabolism of Other Amino Acids
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
E-Flux2 and SPOT
To understand the unknown metabolic mechanism
To validate the predictive power of a metabolic model
To examine the metabolic impacts of inactivation of a gene of interest

Fig 5.1 Summary of this study. We developed computational methods called E-Flux2 and SPOT for inferring system-level and condition-specific intracellular metabolic flux distributions from transcriptomic data (Chapter 2). Their predictive accuracy was validated using the 43 experimental conditions of dataset, which is the largest compiled to date (Chapters 2 and 4). We have shown that these computational tools can aid to understand the unknown metabolic mechanism behind the lipid accumulation mechanism of nitrogen-stressed *P. tricornutum* cells (Chapter 3, section 3.1), validate the predictive power of a newly constructed genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002 (section 3.2), and examine the possible metabolic roles of RpiRc in *S. aureus* (section 3.3).

Appendix

Appendix 1. A mathematical justification for dropping the $\|v\|$ term in the objective function of SPOT

Removing the $\|\boldsymbol{v}\|$ term in the objective function of SPOT can be justified if a solution of problem 2 (see below, on the right side) also optimizes problem 1 (on the left side):

Problem 1

$$\max \frac{v \cdot g}{\|v\|}$$

$$\text{subject to} \begin{cases} Av = 0 \\ 0 \leq v \end{cases}$$

$$\rightarrow$$

Problem 2

$$\max v \cdot g$$

$$\text{subject to} \begin{cases} Av = 0 \\ 0 \leq v \\ \|v\| \leq 1 \end{cases}$$

where v and g has dimension $n,$ and A is a stoichiometric matrix.

Lemma. If v^* optimizes problem 2, then v^* optimizes problem 1.

This lemma can be proved by contradiction.

Let us assume that this statement is false. Then its negation, i.e. if $\boldsymbol{v}^*$ optimizes (i.e. maximizes) problem 2, then $\boldsymbol{v}^*$ does not optimize problem 1, should be true. In other words, we assume that another vector $\boldsymbol{v}$ exists such that

$$\frac{v \cdot g}{\|v\|} > \frac{v^* \cdot g}{\|v^*\|}$$

where $v \neq v^*$.

Let $v' = \frac{v}{\|v\|}$, then $\|v'\| = \frac{1}{\|v\|} \times \|v\| = 1$.

Substituting v' into the term on the left side of the inequality above gives $v' \cdot g$.

In addition, since v^* optimizes problem 2 by assumption, $\|v^*\| \leq 1$. Note that no optimal solution to problem 2 has $\|v\| < 1$ because g is a non-negative vector and we assume that $g \neq 0$ (for otherwise, the problem is trivial). Thus, if $\|v\| < 1$, the objective of problem 2 can always be increased by scaling up v until $\|v\| = 1$. Hence $\|v^*\| = 1$.

The right side of the inequality, thus, is equal to $v^* \cdot g$. Therefore,

$$v' \cdot g > v^* \cdot g$$

where $\|v'\| = 1$.

This contradicts our assumption that v^* is a vector that optimizes problem 2. Since the negation is impossible (false), the original statement is true. Note that the lemma is still valid if the number used to limit $\|v\|$ in problem 2 is any constant $c > 0$.

Appendix 2. A mathematical proof of the uniqueness of SPOT solutions

SPOT (equation (8) in the main manuscript) is defined as the following optimization problem:

$$\max f'v$$

$$\text{subject to} \begin{cases} Av = 0 \\ 0 \leq v \\ \|v\| \leq 1 \end{cases}$$

where v has dimension n, $f \geq 0$ and A is a stoichiometric matrix.

Lemma. If the optimal value of the SPOT problem is strictly positive, then its solution is unique.

Proof. *Step 1.* First we prove that $\|v\| = 1$. If $\|v\| < 1$ then for ϵ sufficiently small $\omega = v + \epsilon \frac{v}{\|v\|}$ satisfies $\|\omega\| < 1$ (and the other constraints) and

$$f^{'}\omega = f^{'}v\left(1 + \frac{\epsilon}{\|v\|}\right) > f'v$$

because $f^{'}v > 0$, contradicting the maximality of $f'v$.

Step 2. Because of the maximality of $f'v$, the affine plane $\{\omega: f^{'}\omega = f'v\}$ is a supporting plane for the convex set $C = \{\omega: \omega \in ker(A),\ \omega \geq 0,\ \|\omega\| \leq 1\}$. Moreover, by Step 1, such a plane cannot be orthogonal to any linear space containing the vector v. We conclude that the affine plane intersects C only at v, thus v is the only point of the maximum for the linear functional $\omega \rightarrow f'\omega$.

Remark. Notice that the condition of strict positivity of the maximum if necessary. Indeed if $ker(A)$ is contained on a coordinate plane, say the first: $ker(A) \subset \{\omega: \omega_1 = 0\}$, then the vector $f = e_1$ satisfies $f^{'}\omega = 0$ for every vector in $ker(A)$ and there is no uniqueness.

Appendix 3. Calculation of the possible range of correlation between the measured fluxes and the predicted fluxes

For standard FBA and E-Flux, the methods that do not give a unique metabolic flux distribution, the possible range of correlation between the measured fluxes and the predicted fluxes were calculated. Given information on the measured fluxes, the minimum and the maximum correlations of standard FBA can be calculated using the following two steps of optimization:

Step 1. standard FBA

$$z^* = \max f'v$$

$$\text{subject to} \begin{cases} Sv = 0 \\ a_j \le v_j \le b_j \end{cases}$$

Step 2. calculation of the possible range of correlation

$$\text{min/max} \, \frac{v_p \cdot v_m}{\|v_p\| \|v_m\|}$$

$$\text{subject to} \begin{cases} Sv = 0 \\ a_j \le v_j \le b_j \\ f^{'}v = z^* \end{cases}$$

where v is a flux vector representing the reaction rates of the n reactions in the network, f is a coefficient vector defining the organism's objective function, S is the stoichiometric matrix, and a_j and b_j are the minimum and maximum reaction rates through reaction j. The vectors v_p and v_m shown in the objective function of Step 2 are the predicted and measured vectors of intracellular fluxes, respectively, and $\|\cdot\|$ denotes the l^2 norm.

The average correlation of standard FBA in Table 2.3 was calculated using solutions obtained in Step 1 under our computational settings (see Methods in the main manuscript for the detailed settings). Step 2 allows us to find a metabolic flux distribution which achieves theoretically maximal or minimal correlation with the measured fluxes while maintaining the optimal biomass flux, denoted as z^* here. The nonlinear optimization problem in Step 2 was solved using the sequential quadratic programming (SQP) algorithm provided by the MATLAB function fmincon.

Importantly, the maximum possible correlation can be calculated only when we already have the known measured flux datasets. There is no way to force each method to produce a metabolic flux distribution which achieves the best correlation with the measured fluxes. Our methods were developed during the process of finding that way and of rigorously testing various strategies.

In the same way, the lower and upper bound of correlations of E-Flux can be calculated as follows:

Step 1. E-Flux

Step 2. calculation of the possible range of correlation

$$z^* = \max f'v \quad \rightarrow \quad \text{min/max} \frac{v_p \cdot v_m}{\|v_p\| \|v_m\|}$$

$$\text{subject to} \begin{cases} Sv = 0 \\ a_j^e \le v_j \le b_j^e \end{cases} \qquad \text{subject to} \begin{cases} Sv = 0 \\ a_j^e \le v_j \le b_j^e \\ f'v = z^* \end{cases}$$

where, for all j,

$$a_j^e = \begin{cases} -g_j, \text{if } a_j < 0, \\ 0, \text{if } a_j \ge 0, \end{cases} \text{ and } b_j^e = \begin{cases} g_j, \text{if } b_j > 0, \\ 0, \text{if } a_j \le 0. \end{cases}$$

Bibliography

1. Stephanopoulos G. Metabolic fluxes and metabolic engineering. Metab Eng. 1999;1: 1–11. doi:10.1006/mben.1998.0101

2. Zelezniak A, Sheridan S, Patil KR. Contribution of Network Connectivity in Determining the Relationship between Gene Expression and Metabolite Concentration Changes. Hatzimanikatis V, editor. PLoS Comput Biol. Public Library of Science; 2014;10: e1003572. doi:10.1371/journal.pcbi.1003572

3. Wiechert W. ^{13}C metabolic flux analysis. Metab Eng. 2001;3: 195–206. doi:10.1006/mben.2001.0187

4. Zamboni N, Fendt S-M, Rühl M, Sauer U. (13)C-based metabolic flux analysis. Nat Protoc. 2009;4: 878–92. doi:10.1038/nprot.2009.58

5. Beurton-Aimar M, Beauvoit B, Monier A, Vallée F, Dieuaide-Noubhani M, Colombié S. Comparison between elementary flux modes analysis and ^{13}C-metabolic fluxes measured in bacterial and plant cells. BMC Syst Biol. 2011;5: 95. doi:10.1186/1752-0509-5-95

6. Nielsen J. It is all about metabolic fluxes. J Bacteriol. 2003;185: 7031–7035. doi:10.1128/JB.185.24.7031-7035.2003

7. Krömer J, Quek L-E, Nielsen L. ^{13}C-fluxomics: A tool for measuring metabolic phenotypes. Aust Biochem. Australian Society of Biochemical and Molecular Biology (ASBMB); 2009;40: 17–20. Available: http://espace.library.uq.edu.au/view/UQ:217071#.U-SiKTwoJWs.mendeley

8. Sauer U. Metabolic networks in motion: ^{13}C-based flux analysis. Mol Syst Biol. 2006;2: 62. doi:10.1038/msb4100109

9. Celton M, Sanchez I, Goelzer A, Fromion V, Camarasa C, Dequin S. A comparative transcriptomic, fluxomic and metabolomic analysis of the response of Saccharomyces cerevisiae to increases in NADPH oxidation. BMC Genomics. 2012. p. 317. doi:10.1186/1471-2164-13-317

10. Winter G, Krömer JO. Fluxomics - connecting ’omics analysis and phenotypes. Environ Microbiol. 2013;15: 1901–16. doi:10.1111/1462-2920.12064

11. Orth JD, Thiele I, Palsson BØ . What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248. doi:10.1038/nbt.1614

12. Blazier AS, Papin JA. Integration of expression data in genome-scale metabolic network reconstructions. Frontiers in Physiology. 2012. doi:10.3389/fphys.2012.00299

13. Chavali AK, D’Auria KM, Hewlett EL, Pearson RD, Papin JA. A metabolic network approach for the identification and prioritization of antimicrobial drug targets. Trends in Microbiology. 2012. pp. 113–123. doi:10.1016/j.tim.2011.12.004

14. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Mol Biosyst. 2013;9: 167–74. doi:10.1039/c2mb25453k

15. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. Nat Rev Genet. 2006;7: 130–141. doi:10.1038/nrg1769

16. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol. 2004;2: 886–897. doi:10.1038/nrmicro1023

17. Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, et al. The transcription unit architecture of the Escherichia coli genome. Nat Biotechnol. 2009;27: 1043–1049. doi:10.1038/nbt.1582

18. Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology. 2010;156: 287–301. doi:10.1099/mic.0.034793-0

19. Palsson B. *In silico* biology through “omics”. Nat Biotechnol. 2002;20: 649–650. doi:10.1038/nbt0702-649

20. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods. Nature Reviews Microbiology. 2012. doi:10.1038/nrmicro2737

21. German JB, Gillies LA, Smilowitz JT, Zivkovic AM, Watkins SM. Lipidomics and lipid profiling in metabolomics. Curr Opin Lipidol. 2007;18: 66–71. doi:10.1097/MOL.0b013e328012d911

22. Mayr M. Metabolomics: ready for the prime time? Circ Cardiovasc Genet. 2008;1: 58–65. doi:10.1161/CIRCGENETICS.108.808329

23. Hoppe A. What mRNA Abundances Can Tell us about Metabolism. Metabolites. 2012. pp. 614–631. doi:10.3390/metabo2030614

24. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484

25. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011;9: 34. doi:10.1186/1741-7007-9-34

26. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol. 2010;6: 787–789. doi:10.1038/nchembio.462

27. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A. 2005;102: 2685–2689. doi:10.1073/pnas.0406811102

28. Bradley PH, Brauer MJ, Rabinowitz JD, Troyanskaya OG. Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. PLoS Comput Biol. 2009;5. doi:10.1371/journal.pcbi.1000270

29. Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, Wheeler RT, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. Proc Natl Acad Sci U S A. 2009;106: 6477–6482. doi:10.1073/pnas.0811091106

30. Reed JL. Shrinking the Metabolic Solution Space Using Experimental Datasets. PLoS Computational Biology. 2012. p. e1002662. doi:10.1371/journal.pcbi.1002662

31. Saha R, Chowdhury A, Maranas CD. Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotechnol. 2014;29C: 39–45. doi:10.1016/j.copbio.2014.02.011

32. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol. 2014;10: e1003580. doi:10.1371/journal.pcbi.1003580

33. Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson B, Hyduke DR. GIM3E: Condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics. 2013;29: 2900–2908. doi:10.1093/bioinformatics/btt493

34. Kim HU, Kim WJ, Lee SY. Flux-coupled genes and their use in metabolic flux analysis. Biotechnol J. 2013;8: 1035–1042. doi:10.1002/biot.201200279

35. Kim J, Reed JL. RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. Genome Biology. 2012. p. R78. doi:10.1186/gb-2012-13-9-r78

36. Collins SB, Reznik E, Segr?? D. Temporal Expression-based Analysis of Metabolism. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002781

37. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2010;107: 17845–17850. doi:10.1073/pnas.1005139107

38. Jensen PA, Papin JA. Functional integration of a metabolic network model and expression data without arbitrary thresholding. Bioinformatics. 2011;27: 541–547. doi:10.1093/bioinformatics/btq702

39. Van Berlo RJP, De Ridder D, Daran JM, Daran-Lapujade PAS, Teusink B, Reinders MJT. Predicting metabolic fluxes using gene expression differences as constraints. IEEE/ACM Trans Comput Biol Bioinforma. 2011;8: 206–216. doi:10.1109/TCBB.2009.55

40. ??kesson M, F??rster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004;6: 285–293. doi:10.1016/j.ymben.2003.12.002

41. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 2003;13: 244–253. doi:10.1101/gr.234503

42. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4: e1000082. doi:10.1371/journal.pcbi.1000082

43. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26: 3140–3142. doi:10.1093/bioinformatics/btq602

44. Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 2008;26: 1003–1010. doi:10.1038/nbt.1487

45. Rossell S, Huynen MA, Notebaart RA. Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1002988

46. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, et al. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009;5: e1000489. doi:10.1371/journal.pcbi.1000489

47. Brandes A, Lun DS, Ip K, Zucker J, Colijn C, Weiner B, et al. Inferring Carbon Sources from Gene Expression Profiles Using Metabolic Flux Models. PLoS ONE. 2012. p. e36947. doi:10.1371/journal.pone.0036947

48. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, et al. Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol. 2012;6: 73. doi:10.1186/1752-0509-6-73

49. Reder C. Metabolic control theory: a structural approach. J Theor Biol. 1988;135: 175–201.

50. Barker B, Sadagopan N, Wang Y, Smallbone K, Myers CR, Xi H, et al. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. 2014; 27. Available: http://arxiv.org/abs/1404.4755

51. Pramanik J, Keasling JD. Stoichiometric model of Escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng. 1997;56: 398–421. doi:10.1002/(SICI)1097-0290(19971120)56:4<398::AID-BIT6>3.0.CO;2-J

52. Westerhoff H V, Hellingwerf KJ, Van Dam K. Thermodynamic efficiency of microbial growth is low but optimal for maximal growth rate. Proc Natl Acad Sci U S A. 1983;80: 305–309. doi:10.1073/pnas.80.1.305

53. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010;13: 344–349. doi:10.1016/j.mib.2010.03.003

54. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Brief Bioinform. 2009;10: 435–449. doi:10.1093/bib/bbp011

55. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5: 621–628. doi:10.1038/nmeth.1226

56. Cocaign M, Monnet C, Lindley N. Batch kinetics of Corynebacterium glutamicum during growth on various carbon substrates: use of substrate mixtures to localise metabolic bottlenecks. Applied Microbiology and Biotechnology. 1993. doi:10.1007/BF00175743

57. Kim MK, Lun DS. Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J. 2014;11: 59–65. doi:10.1016/j.csbj.2014.08.009

58. Joyce A, Palsson B, Joyce AR, Palsson BØ . The model organism as a system: integrating “omics” data sets. Nat Rev Mol Cell Biol. 2006;7: 198–210. doi:10.1038/nrm1857

59. Song H-S, Reifman J, Wallqvist A. Prediction of metabolic flux distribution from gene expression data based on the flux minimization principle. PLoS One. 2014;9: e112524. doi:10.1371/journal.pone.0112524

60. Smallbone K, Simeonidis E. Flux balance analysis: A geometric perspective. J Theor Biol. 2009;258: 311–315. doi:10.1016/j.jtbi.2009.01.027

61. Dauner M, Sonderegger M, Hochuli M, Szyperski T, Wüthrich K, Hohmann H-P, et al. Intracellular carbon fluxes in riboflavin-producing Bacillus subtilis during growth on two-carbon substrate mixtures. Appl Environ Microbiol. 2002;68: 1760–1771. doi:10.1128/AEM.68.4.1760-1771.2002

62. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, et al. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. Science. 2007;316: 593–597. doi:10.1126/science.1132067

63. Holm AK, Blank LM, Oldiges M, Schmid A, Solem C, Jensen PR, et al. Metabolic and transcriptional response to cofactor perturbations in *Escherichia coli*. J Biol Chem. 2010;285: 17498–17506. doi:10.1074/jbc.M109.095570

64. Rintala E, Toivari M, Pitkänen J-P, Wiebe MG, Ruohonen L, Penttilä M. Low oxygen levels as a trigger for enhancement of respiratory metabolism in Saccharomyces cerevisiae. BMC Genomics. 2009;10: 461. doi:10.1186/1471-2164-10-461

65. Jouhten P, Rintala E, Huuskonen A, Tamminen A, Toivari M, Wiebe M, et al. Oxygen dependence of metabolic fluxes and energy generation of Saccharomyces cerevisiae CEN.PK113-1A. BMC Syst Biol. 2008;2: 60. doi:10.1186/1752-0509-2-60

66. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Molecular Systems Biology. 2011. doi:10.1038/msb.2011.65

67. Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP. Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. BMC Syst Biol. 2012;6: 55. doi:10.1186/1752-0509-6-55

68. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007;3: 121. doi:10.1038/msb4100155

69. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 2003;4: R54. doi:10.1186/gb-2003-4-9-r54

70. Mo ML, Palsson BO, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol. 2009;3: 37. doi:10.1186/1752-0509-3-37

71. Duarte NC, Herrgård MJ, Palsson BØ . Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res. 2004;14: 1298–1309. doi:10.1101/gr.2250904

72. Wu H, Kerr M, Cui X, Churchill G, Yang H. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. … data: methods and software. 2003. pp. 313–341. doi:doi:10.1007/0-387-21679-0_14

73. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4: 249–264. doi:10.1093/biostatistics/4.2.249

74. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, et al. A new

non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002;3: research0048. doi:10.1186/gb-2002-3-9-research0048

75. Dudley AM, Aach J, Steffen M a, Church GM. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci U S A. 2002;99: 7554–7559. doi:10.1073/pnas.112683499

76. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol. 2006;195: 373–388. doi:10.1016/j.ajog.2006.07.001

77. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics. 2006. pp. 101–109. doi:10.1016/j.tig.2005.12.005

78. Reimers M. Making informed choices about microarray data analysis. PLoS Comput Biol. 2010;6: 1–7. doi:10.1371/journal.pcbi.1000786

79. Bonarius HP, Hatzimanikatis V, Meesters KP, de Gooijer CD, Schmid G, Tramper J. Metabolic flux analysis of hybridoma cells in different culture media using mass balances. Biotechnol Bioeng. 1996;50: 299–318. doi:10.1002/(SICI)1097-0290(19960505)50:3<299::AID-BIT9>3.0.CO;2-B

80. Bertsekas D, Nedić A, Ozdaglar A. Convex Analysis and Optimization. Athena Scientific; 2003.

81. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6: 390. doi:10.1038/msb.2010.47

82. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. Crit Care. 2003;7: 451–459. doi:10.1186/cc2401

83. Ray WJ. Rate-limiting step: a quantitative definition. Application to steady-state enzymic reactions. Biochemistry. 1983;22: 4625–4637. doi:10.1021/bi00289a003

84. Keating SM, Bornstein BJ, Finney A, Hucka M. SBMLToolbox: an SBML toolbox for MATLAB users. Bioinformatics. 2006;22: 1275–1277. doi:10.1093/bioinformatics/btl111

85. Kelley JJ, Lane A, Li X, Mutthoju B, Maor S, Egen D, et al. MOST: a software environment for constraint-based metabolic modeling and strain design. Bioinforma . 2014; doi:10.1093/bioinformatics/btu685

86. Taylor R. Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography. 1990. pp. 35–39.

doi:10.1177/875647939000600106

87. Klitgord N, Segrè D. The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. Genome Inform. 2010;22: 41–55. doi:10.1142/9781848165786_0005

88. Hyduke D, Hyduke D, Schellenberger J, Que R, Fleming R, Thiele I, et al. COBRA Toolbox 2.0. Protoc Exch. 2011; 1–35. doi:10.1038/protex.2011.234

89. Nevoigt E. Progress in metabolic engineering of Saccharomyces cerevisiae. Microbiol Mol Biol Rev. 2008;72: 379–412. doi:10.1128/MMBR.00025-07

90. Görke B, Stülke J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nat Rev Microbiol. 2008;6: 613–624. doi:10.1038/nrmicro1932

91. Vinuselvi P, Kim MK, Lee SK, Ghim CM. Rewiring carbon catabolite repression for microbial cell factory. BMB Reports. 2012. pp. 59–70. doi:10.5483/BMBRep.2012.45.2.59

92. Levitan O, Dinamarca J, Zelzion E, Lun DS, Guerra LT, Kim MK, et al. Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress. Proc Natl Acad Sci. 2015;112: 412–417. doi:10.1073/pnas.1419818112

93. Qian X, Kim MK, Kumaraswamy GK, Agarwal A, Lun DS, Dismukes GC. Flux balance analysis of photoautotrophic metabolism: Uncovering new biological details of subsystems involved in cyanobacterial photosynthesis. Biochim Biophys Acta - Bioenerg. 2017;1858: 276–287. doi:10.1016/j.bbabio.2016.12.007

94. Balasubramanian D, Ohneck EA, Chapman J, Weiss A, Kim MK, Reyes-Robles T, et al. *Staphylococcus aureus* coordinates leukocidin expression and pathogenesis by sensing metabolic fluxes via RpiRc. MBio. 2016;7. doi:10.1128/mBio.00818-16

95. Chisti Y. Biodiesel from microalgae beats bioethanol. Trends Biotechnol. 2008;26: 126–131. doi:10.1016/j.tibtech.2007.12.002

96. Granum E, Raven J a, Leegood RC. How do marine diatoms fix 10 billion tonnes of inorganic carbon per year? Can J Bot. 2005;83: 898–908. doi:10.1139/b05-077

97. Valenzuela J, Mazurie A, Carlson RP, Gerlach R, Cooksey KE, Peyton BM, et al. Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum. Biotechnol Biofuels. 2012;5: 40. doi:10.1186/1754-6834-5-40

98. Merchant SS, Kropat J, Liu B, Shaw J, Warakanont J. TAG, You're it! Chlamydomonas as a reference organism for understanding algal triacylglycerol

accumulation. Current Opinion in Biotechnology. 2012. pp. 352–363. doi:10.1016/j.copbio.2011.12.001

99. Korman TP, Sahachartsiri B, Charbonneau DM, Huang GL, Beauregard M, Bowie JU. Dieselzymes: development of a stable and methanol tolerant lipase for biodiesel production by directed evolution. Biotechnol Biofuels. 2013;6: 70. doi:10.1186/1754-6834-6-70

100. Guerra LT, Levitan O, Frada MJ, Sun JS, Falkowski PG, Dismukes GC. Regulatory branch points affecting protein and lipid biosynthesis in the diatom Phaeodactylum tricornutum. Biomass and Bioenergy. 2013;59: 306–315. doi:10.1016/j.biombioe.2013.10.007

101. Goldman JC, Mccarthy JJ. Steady state growth and ammonium uptake of a fast-growing marine diatom. Limnol Oceanogr. 1978;23: 695–703. doi:10.4319/lo.1978.23.4.0695

102. Guillard RRL, Ryther JH. Studies of marine planktonic diatoms: I. Cyclotella nana Hustedt, and Detonula confervacea (Cleve) Gran. Can J Microbiol. 1962;8: 229–239. doi:10.1139/m62-029

103. Kim J, Fabris M, Baart G, Kim MK, Goossens A, Vyverman W, et al. Flux balance analysis of primary metabolism in the diatom Phaeodactylum tricornutum. Plant J. 2015;85: n/a-n/a. doi:10.1111/tpj.13081

104. Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE. The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. Plant J. 2012;70: 1004–1014. doi:10.1111/j.1365-313X.2012.04941.x

105. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40. doi:10.1093/nar/gkr988

106. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016;44: D471–D480. doi:10.1093/nar/gkv1164

107. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, et al. BRENDA in 2013: Integrated reactions, kinetic data, enzyme function data, improved disease classification: New options and contents in BRENDA. Nucleic Acids Res. 2013;41. doi:10.1093/nar/gks1049

108. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–10. doi:10.1016/S0022-2836(05)80360-2

109. Anders S, Huber W. Differential expression analysis for sequence count data.

Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

110. De Martino A, Bartual A, Willis A, Meichenin A, Villazán B, Maheswari U, et al. Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom Phaeodactylum tricornutum. Protist. 2011;162: 462–481. doi:10.1016/j.protis.2011.02.002

111. Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. New Phytol. 2010;188: 52–66. doi:10.1111/j.1469-8137.2010.03371.x

112. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

113. Kim MK, Lane A, Kelley JJ, Lun DS. E-Flux2 and SPOT: Validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. PLoS One. 2016;11. doi:10.1371/journal.pone.0157101

114. Frada MJ, Burrows EH, Wyman KD, Falkowski PG. Quantum requirements for growth and fatty acid biosynthesis in the marine diatom Phaeodactylum tricornutum (Bacillariophyceae) in nitrogen replete and limited conditions. J Phycol. 2013;49: 381–388. doi:10.1111/jpy.12046

115. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics. 2011;27: 431–432. doi:10.1093/bioinformatics/btq675

116. Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, et al. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. TL - 473. Nature. 2011;473 VN-: 203–207. doi:10.1038/nature10074

117. Mailloux RJ, Bériault R, Lemire J, Singh R, Chénier DR, Hamel RD, et al. The tricarboxylic acid cycle, an ancient metabolic network with a novel twist. PLoS One. 2007;2. doi:10.1371/journal.pone.0000690

118. Yang Z-K, Niu Y-F, Ma Y-H, Xue J, Zhang M-H, Yang W-D, et al. Molecular and cellular mechanisms of neutral lipid accumulation in diatom following nitrogen deprivation. Biotechnol Biofuels. 2013;6: 67. doi:10.1186/1754-6834-6-67

119. Levitan O, Dinamarca J, Zelzion E, Gorbunov MY, Falkowski PG. An RNA interference knock-down of nitrate reductase enhances lipid biosynthesis in the diatom Phaeodactylum tricornutum. Plant J. 2015;84: 963–973. doi:10.1111/tpj.13052

120. Parmar A, Singh NK, Pandey A, Gnansounou E, Madamwar D. Cyanobacteria and microalgae: A positive prospect for biofuels. Bioresource Technology. 2011. pp. 10163–10172. doi:10.1016/j.biortech.2011.08.030

121. Quintana N, Van Der Kooy F, Van De Rhee MD, Voshol GP, Verpoorte R. Renewable energy from Cyanobacteria: Energy production optimization by metabolic pathway engineering. Applied Microbiology and Biotechnology. 2011. pp. 471–490. doi:10.1007/s00253-011-3394-0

122. Qi F, Yao L, Tan X, Lu X. Construction, characterization and application of molecular tools for metabolic engineering of Synechocystis sp. Biotechnol Lett. 2013;35: 1655–1661. doi:10.1007/s10529-013-1252-0

123. Qian X, Kumaraswamy GK, Zhang S, Gates C, Ananyev GM, Bryant DA, et al. Inactivation of nitrate reductase alters metabolic branching of carbohydrate fermentation in the cyanobacterium *Synechococcus* sp. strain PCC 7002. Biotechnol Bioeng. 2016;113: 979–988. doi:10.1002/bit.25862

124. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC. Genetic engineering of algae for enhanced biofuel production. Eukaryotic Cell. 2010. pp. 486–501. doi:10.1128/EC.00364-09

125. McNeely K, Xu Y, Bennette N, Bryant DA, Dismukes GC. Redirecting reductant flux into hydrogen production via metabolic engineering of fermentative carbon metabolism in a cyanobacterium. Appl Environ Microbiol. 2010;76: 5032–5038. doi:10.1128/AEM.00862-10

126. Zhang S, Bryant DA. The tricarboxylic acid cycle in cyanobacteria. Science. 2011;334: 1551–3. doi:10.1126/science.1210858

127. Liu X, Fallon S, Sheng J, Curtiss R. CO_2-limitation-inducible Green Recovery of fatty acids from cyanobacterial biomass. Proc Natl Acad Sci U S A. 2011;108: 6905–8. doi:10.1073/pnas.1103016108

128. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ . Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009;7: 129–43. doi:10.1038/nrmicro1949

129. Knoop H, Zilliges Y, Lockau W, Steuer R. The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol. 2010;154: 410–22. doi:10.1104/pp.110.157198

130. Knoop H, Gründel M, Zilliges Y, Lehmann R, Hoffmann S, Lockau W, et al. Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of Synechocystis sp. PCC 6803. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1003081

131. Hendry JI, Prasannan CB, Joshi A, Dasgupta S, Wangikar PP. Metabolic model of *Synechococcus* sp. PCC 7002: Prediction of flux distribution and network modification for enhanced biofuel production. Bioresour Technol. 2016;213: 190–197. doi:10.1016/j.biortech.2016.02.128

132. Hamilton JJ, Reed JL. Identification of functional differences in metabolic networks using comparative genomics and constraint-based models. PLoS One. 2012;7. doi:10.1371/journal.pone.0034670

133. Vu TT, Hill EA, Kucek LA, Konopka AE, Beliaev AS, Reed JL. Computational evaluation of *Synechococcus* sp. PCC 7002 metabolism for chemical production. Biotechnol J. 2013;8: 619–630. doi:10.1002/biot.201200315

134. Frigaard, Niels-Ulrik; Sakuragi, Yumiko; Bryant DA. Gene Inactivation in the Cyanobacterium *Synechococcus* sp. PCC 7002 and the Green Sulfur Bacterium Chlorobium tepidum Using In Vitro-Made DNA Constructs and Natural Transformation. Methods in molecular biology (Clifton, N.J.). 2004. doi:10.1385/1-59259-799-8

135. Xu Y, Alvey RM, Byrne PO, Graham JE, Shen G, Bryant DA. Expression of genes in cyanobacteria: adaptation of endogenous plasmids as platforms for high-level gene expression in *Synechococcus* sp. PCC 7002. Methods Mol Biol. 2011;684: 273–93. doi:10.1007/978-1-60761-925-3_21

136. Cornet J-F, Dussap C, Gros J-B. Kinetics and Energetics of Photosynthetic Micro-Organisms in Photobioreactors Application to Spirulina Growth. Bioprocess Algae React Technol Apoptosis. 1998;59: 153–224. doi:10.1007/BFb0102299

137. Olguín EJ, Galicia S, Angulo-Guerrero O, Hernández E. The effect of low light flux and nitrogen deficiency on the chemical composition of Spirulina sp. (Arthrospira) grown on digested pig waste. Bioresour Technol. 2001;77: 19–24. doi:10.1016/S0960-8524(00)00142-5

138. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010;28: 977–982. doi:10.1038/nbt.1672

139. Heavner BD, Price ND. Transparency in metabolic network reconstruction enables scalable biological discovery. Current Opinion in Biotechnology. 2015. pp. 105–109. doi:10.1016/j.copbio.2014.12.010

140. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. Nat Biotechnol. 2014;32: 447–452. doi:10.1038/nbt.2870

141. Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: The need for a unified standard. Brief Bioinform. 2015;16: 1057–1068. doi:10.1093/bib/bbv003

142. Ludwig M, Bryant DA. Acclimation of the global transcriptome of the cyanobacterium *Synechococcus* sp. strain PCC 7002 to nutrient limitations and different nitrogen sources. Front Microbiol. 2012;3. doi:10.3389/fmicb.2012.00145

143. Hasunuma T, Kikuyama F, Matsuda M, Aikawa S, Izumi Y, Kondo A. Dynamic metabolic profiling of cyanobacterial glycogen biosynthesis under conditions of nitrate depletion. J Exp Bot. 2013;64: 2943–2954. doi:10.1093/jxb/ert134

144. Nigam S, Rai MP, Sharma R. Effect of nitrogen on growth and lipid content of Chlorella pyrenoidosa. Am J Biochem Biotechnol. 2011;7: 126–131.

145. Taikhao S, Incharoensakdi A, Phunpruch S. Dark fermentative hydrogen production by the unicellular halotolerant cyanobacterium Aphanothece halophytica grown in seawater. Journal of Applied Phycology. 2014. doi:10.1007/s10811-014-0292-8

146. Ludwig M, Bryant DA. *Synechococcus* sp. Strain PCC 7002 Transcriptome: Acclimation to Temperature, Salinity, Oxidative Stress, and Mixotrophic Growth Conditions. Frontiers in Microbiology. 2012. doi:10.3389/fmicb.2012.00354

147. Sauer J, Schreiber U, Schmid R, Völker U, Forchhammer K. Nitrogen starvation-induced chlorosis in *Synechococcus* PCC 7942. Low-level photosynthesis as a mechanism of long-term survival. Plant Physiol. 2001;126: 233–243. doi:10.1104/pp.126.1.233

148. Stevens SE, Balkwill DL, Paone DAM. The effects of nitrogen limitation on the ultrastructure of the cyanobacterium *Agmenellum quadruplicatum*. Arch Microbiol. 1981;130: 204–212. doi:10.1007/BF00459520

149. Duke CS, Allen MM. Effect of Nitrogen Starvation on Polypeptide Composition, Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase, and Thylakoid Carotenoprotein Content of *Synechocystis* sp. Strain PCC6308. PLANT Physiol. 1990;94: 752–759. doi:10.1104/pp.94.2.752

150. Griffiths MJ, Harrison STL. Lipid productivity as a key characteristic for choosing algal species for biodiesel production. J Appl Phycol. 2009;21: 493–507. doi:10.1007/s10811-008-9392-7

151. Tedesco MA, Duerr EO. Light, temperature and nitrogen starvation effects on the total lipid and fatty acid content and composition of *Spirulina platensis* UTEX 1928. J Appl Phycol. 1989;1: 201–209. doi:10.1007/BF00003646

152. Guerra LT, Xu Y, Bennette N, McNeely K, Bryant DA, Dismukes GC. Natural osmolytes are much less effective substrates than glycogen for catabolic energy production in the marine cyanobacterium *Synechococcus* sp. strain PCC 7002. J Biotechnol. 2013;166: 65–75. doi:10.1016/j.jbiotec.2013.04.005

153. Vandenesch F, Lina G, Henry T. *Staphylococcus aureus* hemolysins, bi-component leukocidins, and cytolytic peptides: a redundant arsenal of membrane-damaging virulence factors? Front Cell Infect Microbiol. 2012;2: 12. doi:10.3389/fcimb.2012.00012

154. Otto M. *Staphylococcus aureus* toxins. Current Opinion in Microbiology. 2014. pp. 32–37. doi:10.1016/j.mib.2013.11.004

155. Alonzo III F, Torres VJ. The Bicomponent Pore-Forming Leucocidins of *Staphylococcus aureus*. Microbiol Mol Biol Rev. 2014;78: 199–230. doi:10.1128/MMBR.00055-13

156. Menestrina G, Dalla Serra M, Comai M, Coraiola M, Viero G, Werner S, et al. Ion channels and bacterial infection: The case of β-barrel pore-forming protein toxins of *Staphylococcus aureus*. FEBS Letters. 2003. pp. 54–60. doi:10.1016/S0014-5793(03)00850-0

157. Novick RP, Geisinger E. Quorum sensing in staphylococci. AnnuRevGenet. 2008;42: 541–564. doi:10.1146/annurev.genet.42.110807.091640

158. Regassa LB, Betley MJ. Alkaline pH decreases expression of the accessory gene regulator (agr) in *Staphylococcus aureus*. J Bacteriol. 1992;174: 5095–5100.

159. Weinrick B, Dunman PM, McAleese F, Murphy E, Projan SJ, Fang Y, et al. Effect of mild acid on gene expression in *Staphylococcus aureus*. J Bacteriol. 2004;186: 8407–8423. doi:10.1128/JB.186.24.8407-8423.2004

160. Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N, Possedko M, et al. *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. Genes Dev. 2007;21: 1353–1366. doi:10.1101/gad.423507

161. Benson MA, Ohneck EA, Ryan C, Alonzo F, Smith H, Narechania A, et al. Evolution of hypervirulence by a MRSA clone through acquisition of a transposable element. Mol Microbiol. 2014;93: 664–681. doi:10.1111/mmi.12682

162. Killikelly A, Benson MA, Ohneck EA, Sampson JM, Jakoncic J, Spurrier B, et al. Structure-based functional characterization of repressor of toxin (Rot), a central regulator of *Staphylococcus aureus* virulence. J Bacteriol. 2015;197: 188–200. doi:10.1128/JB.02317-14

163. Boles BR, Thoendel M, Roth AJ, Horswill AR. Identification of genes involved in polysaccharide-independent *Staphylococcus aureus* biofilm formation. PLoS One. 2010;5: e10146. doi:10.1371/journal.pone.0010146

164. Chen J, Yoong P, Ram G, Torres VJ, Novick RP. Single-copy vectors for integration at the SaPI1 attachment site for *Staphylococcus aureus*. Plasmid. 2014;76: 1–7. doi:10.1016/j.plasmid.2014.08.001

165. Carroll RK, Weiss A, Shaw LN. RNA-Sequencing of *Staphylococcus aureus* messenger RNA. Methods in Molecular Biology. 2016. pp. 131–141. doi:10.1007/7651_2014_192

166. McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, et al. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. 2013;41. doi:10.1093/nar/gkt444

167. Becker SA, Palsson BØ . Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. BMC Microbiol. 2005;5: 8. doi:10.1186/1471-2180-5-8

168. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, Bose JL, et al. A genetic resource for rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus* genes. MBio. 2013;4. doi:10.1128/mBio.00537-12

169. Sørensen KI, Hove-Jensen B. Ribose catabolism of Escherichia coli: Characterization of the rpiB gene encoding ribose phosphate isomerase B and of the rpiR gene, which is involved in regulation of rpiB expression. J Bacteriol. 1996;178: 1003–1011.

170. Daddaoua A, Krell T, Ramos JL. Regulation of glucose metabolism in Pseudomonas. The phosphorylative branch and Entner-Doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain. J Biol Chem. 2009;284: 21360–21368. doi:10.1074/jbc.M109.014555

171. Yamamoto H, Serizawa M, Thompson J, Sekiguchi J. Regulation of the glv operon in Bacillus subtilis: YfiA (GlvR) is a positive regulator of the operon that is repressed through CcpA and cre. J Bacteriol. 2001;183: 5110–5121. doi:10.1128/JB.183.17.5110-5121.2001

172. Zhu Y, Nandakumar R, Sadykov MR, Madayiputhiya N, Luong TT, Gaupp R, et al. RpiR homologues may link *Staphylococcus aureus* RNAIII synthesis and pentose phosphate pathway regulation. J Bacteriol. 2011;193: 6187–6196. doi:10.1128/JB.05930-11

173. Somerville GA, Proctor RA. At the crossroads of bacterial metabolism and virulence factor synthesis in staphylococci. MicrobiolMolBiolRev. 2009;73: 233–248. doi:10.1128/MMBR.00005-09

174. Ledala N, Zhang B, Seravalli J, Powers R, Somerville GA. Influence of iron and aeration on *Staphylococcus aureus* growth, metabolism, and transcription. J Bacteriol. 2014;196: 2178–2189. doi:10.1128/JB.01475-14

175. Seidl K, Müller S, François P, Kriebitzsch C, Schrenzel J, Engelmann S, et al. Effect of a glucose impulse on the CcpA regulon in *Staphylococcus aureus*. BMC Microbiol. 2009;9: 95. doi:10.1186/1471-2180-9-95

176. Ding Y, Liu X, Chen F, Di H, Xu B, Zhou L, et al. Metabolic sensor governing bacterial virulence in *Staphylococcus aureus*. Proc Natl Acad Sci. 2014;111: E4981–E4990. doi:10.1073/pnas.1411077111

177. Pohl K, Francois P, Stenz L, Schlink F, Geiger T, Herbert S, et al. CodY in *Staphylococcus aureus*: A regulatory link between metabolism and virulence gene expression. J Bacteriol. 2009;191: 2953–2963. doi:10.1128/JB.01492-08

178. Sonenshein AL. CodY, a global regulator of stationary phase and virulence in Gram-positive bacteria. Current Opinion in Microbiology. 2005. pp. 203–207. doi:10.1016/j.mib.2005.01.001

179. Peregrín-Alvarez JM, Sanford C, Parkinson J. The conservation and evolutionary modularity of metabolism. Genome Biol. 2009;10: R63. doi:10.1186/gb-2009-10-6-r63

180. Rodrigues F, Ludovico P, Leão C. Sugar Metabolism in Yeasts : an Overview of Aerobic and Anaerobic Glucose Catabolism. Biodivers Ecophysiol Yeasts. 2006; 101–121. doi:10.1007/3-540-30985-3_6

181. Uzman A, Lodish H, Berk A, Zipursky L, Baltimore D. Molecular Cell Biology (4th edition) New York, NY, 2000, ISBN 0-7167-3136-3. Biochem Mol Biol Educ. 2000;29: Section 1.2The Molecules of Life. doi:10.1016/S1470-8175(01)00023-6

182. Bassham JA, Krause GH. Free energy changes and metabolic regulation in steady-state photosynthetic carbon reduction. Biochim Biophys Acta. 1969;45: 207–221. doi:10.1016/0005-2728(69)90048-6

183. Gerosa L, Haverkorn Van Rijsewijk BRB, Christodoulou D, Kochanowski K, Schmidt TSB, Noor E, et al. Pseudo-transition Analysis Identifies the Key Regulators of Dynamic Metabolic Adaptations from Steady-State Data. Cell Syst. 2015;1: 270–282. doi:10.1016/j.cels.2015.09.008

184. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. Mol Syst Biol. 2011;7: 535. doi:10.1038/msb.2011.65

185. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012;335: 1103–6. doi:10.1126/science.1206848

186. Chubukov V, Uhr M, Le Chat L, Kleijn RJ, Jules M, Link H, et al. Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis. Mol Syst Biol. 2013;9: 709. doi:10.1038/msb.2013.66

187. Oh Y-K, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem. American Society for Biochemistry and Molecular Biology; 2007;282: 28791–9. doi:10.1074/jbc.M703759200

188. You L, He L, Tang YJ. Photoheterotrophic fluxome in Synechocystis sp. strain

PCC 6803 and its implications for cyanobacterial bioenergetics. J Bacteriol. 2015;197: 943–950. doi:10.1128/JB.02149-14

189. Young JD, Shastri AA, Stephanopoulos G, Morgan JA. Mapping photoautotrophic metabolism with isotopically nonstationary ^{13}C flux analysis. Metab Eng. 2011;13: 656–665. doi:10.1016/j.ymben.2011.08.002

190. You L, Berla B, He L, Pakrasi HB, Tang YJ. ^{13}C-MFA delineates the photomixotrophic metabolism of Synechocystis sp. PCC 6803 under light- and carbon-sufficient conditions. Biotechnol J. 2014;9: 684–692. doi:10.1002/biot.201300477

191. Sriram G, Parr LS, Rahib L, Liao JC, Dipple KM. Moonlighting function of glycerol kinase causes systems-level changes in rat hepatoma cells. Metab Eng. 2010;12: 332–340. doi:10.1016/j.ymben.2010.04.001

192. Sriram G, Rahib L, He J Sen, Campos AE, Parr LS, Liao JC, et al. Global metabolic effects of glycerol kinase overexpression in rat hepatoma cells. Mol Genet Metab. 2008;93: 145–159. doi:10.1016/j.ymgme.2007.09.008

193. Gille C, Bölling C, Hoppe A, Bulik S, Hoffmann S, Hübner K, et al. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol. 2010;6: 411. doi:10.1038/msb.2010.62

194. Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BØ. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol. 2010;4: 140. doi:10.1186/1752-0509-4-140

195. Persson J, Fink P, Goto A, Hood JM, Jonas J, Kato S. To be or not to be what you eat: Regulation of stoichiometric homeostasis among autotrophs and heterotrophs. Oikos. 2010;119: 741–751. doi:10.1111/j.1600-0706.2009.18545.x

196. Sterner RW, Clasen J, Lampert W, Weisse T. Carbon:phosphorus stoichiometry and food chain production. Ecol Lett. 1998;1: 146–150. doi:10.1046/j.1461-0248.1998.00030.x

Curriculum Vitae

Min Kyung Kim
mk1034@rutgers.edu
+1-856-671-1815

Education

Ph.D. in Computational and Integrative Biology, **Rutgers University**, May 2017
M.S. in Biomedical Sciences, **Seoul National University**, Feb 2011
B.S. in Biotechnology and **B.A.** in International Studies, **Korea University**, Feb 2009

Research Experience

2012.9-present Graduate assistant, Center for Computational and Integrative Biology at Rutgers University (Advisor: Prof. Desmond S. Lun)

- Developed computational tools for inferring system-level and condition-specific intracellular metabolic flux distribution from gene expression data
- **Dissertation:** Inference of metabolic flux distributions from transcriptomic data

2011-2012 Post-master researcher, Laboratory of Quantitative Biology and Biophysics, Department of Nano-biochemical Engineering at Ulsan National Institute of Science and Technology (UNIST) (Advisor: Prof. Cheol-Min Ghim)

- Worked on constructing a genome-scale metabolic model of *Zymomonas mobilis*

2009-2011 Research assistant, Laboratory of neurobiology for mental disorders, Department of Biomedical Sciences at Seoul National University, Seoul National University Hospital Biomedical Institute (Advisor: Prof. Yong-Sik Kim)

- Studied the effect of antipsychotics on cellular signaling pathways in rat brain and liver using biochemical techniques
- **Thesis:** The effect of clozapine on the AMPK-ACC-CPT1 pathway in the rat frontal cortex

2008 Undergraduate researcher, Laboratory of Virus-Host Interactions, Department of Biotechnology at Korea University (Advisor: Prof. Moon-Jung Song)

- Worked on constructing ORF11 STOP viral plasmid of Murine Gammaherpesvirus 68 by allelic exchange
- **Thesis:** Screening of Bacterial Artificial Chromosome-based ORF11 STOP recombinant viral genome of Murine gammaherpesvirus 68

Publications

*Google Scholar link:
https://scholar.google.com/citations?hl=en&user=x138sVQAAAAJ&view_op=list_works&sortby=pubdate

1. James J. Kelly, Shay Maor, Min Kyung Kim, Anatoliy Lane, Desmond S. Lun, **MOST-Visualization: Software for producing automated textbook-style maps of genome-scale metabolic networks**, Bioinformatics, doi: 10.1093/bioinformatics/btx240

2. Xiao Qian, Min Kyung Kim, G. Kenchappa Kumaraswamy, Ananya Agarwal, Desmond S. Lun, and G. Charles Dismukes, **Flux balance analysis of photoautotrophic**

metabolism: Uncovering new biological details of subsystems involved in cyanobacterial photosynthesis, *BBA-Bioenergetics*, 2017, 1858: 276–287

3. Nick Fyson, Min Kyung Kim, Desmond S. Lun, Caroline Colijn, Gene-centric constraint of metabolic models, *bioRxiv*, 2017, 116558

4. Min Kyung Kim, Anatoliy Lane, James J. Kelly, and Desmond S. Lun, **E-Flux2 and SPOT: Validated methods for inferring intracellular metabolic flux distributions from transcriptomic data**, *PLOS One*, 2016, 11 (6), e0157101

5. Divya Balasubramanian, Elizabeth A. Ohneck, Jessica Lim-Chapman, Andy Weiss, Min Kyung Kim, Tamara Reyes-Robles, Lindsey Shaw, Desmond S. Lun, Beatrix Ueberheide, Bo Shopsin, and Victor J. Torres, ***Staphylococcus aureus* coordinates leukocidins expression and pathogenesis by sensing metabolic fluxes via RpiRC**, *mBio,* 2016, 7 (3), e00818-16

6. Joomi Kim, Michele Fabris, Gino Baart, Min Kyung Kim, Alain Goossens, Wim Vyverman, Paul Falkowski, and Desmond Lun, **Flux balance analysis of primary metabolism in the diatom *Phaeodactylum tricornutum***, *The Plant Journal*, 2016, 85(1):161–176

7. Orly Levitan, Jorge Dinamarca, Ehud Zelzion, Desmond S Lun, L Tiago Guerra, Min Kyung Kim, Joomi Kim, Benjamin AS Van Mooy, Debashish Bhattacharya, and Paul G Falkowski, **Remodeling of intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress,** *PNAS*, 2015 112 (2): 412-417

8. Kuhn Ip, Neil Donoghue, Min Kyung Kim, and Desmond S Lun, **Constraint-based modeling of heterologous pathways: Application and experimental demonstration for overproduction of fatty acids in *Escherichia coli***, *Biotechnology and bioengineering*, 2014 111 (10) : 2056-2066

9. Min Kyung Kim and Desmond S Lun, **Methods for integration of transcriptomic data in genome-scale metabolic models**, *Computational and structural biotechnology journal*, 2014 11 (18): 59-65

10. Min Kyung Kim, Se Hyun Kim, Hyun Sook Yu, Hong Geun Park, Ung Gu Kang, Yong Min Ahn, and Yong Sik Kim, **The effect of clozapine on the AMPK-ACC-CPT1 pathway in the rat frontal cortex**, *International Journal of Neuropsychopharmacology*, 2012 15 (7): 907-917

11. Parisutham Vinuselvi, Min Kyung Kim, Sung Kuk Lee, and Cheol-Min Ghim **Rewiring carbon catabolite repression for microbial cell factory**, *Biochemistry and Molecular Biology Reports*, 2012 45 (2): 59-70

12. Se Hyun Kim, Min Kyung Kim, Hyun Sook Yu, Han Soo Kim, In Sun Park, Hong Geun Park, Ung Gu Kang, and Yong Sik Kim, **Electroconvulsive seizure increases phosphorylation of PKC substrates, including GAP-43, MARCKS, and neurogranin, in rat brain**, *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 2010 34 (1):115-121

Skills and Techniques

1. Computer programming languages: MATLAB, Python, Java, R, C and HTML

2. Laboratory techniques in Biochemistry and Molecular Biology
   : rat brain dissection, rat transcardial perfusion, intraperitoneal injection, intracerebroventricular injection, SH-SY5Y (human neuroblastoma cell line) culture, primary cortical neuronal culture, gel electrophoresis, Western blotting, RNA prep, cDNA synthesis, realtime PCR, gDNA prep, immuohistochemistry, immunocytochemistry, immunofluorescence analysis, cellular fractionation, immunoprecipitation, flow cytometry, carnitine palmitoyl transferase 1 (CPT1) activity assay, blood hormone assay using ELISA, blood glucose level assay, locomotor activity test, pre-pulse inhibition, culturing *Daphna Magna* (water fleas), allelic exchange, colony PCR, culturing bacteria in/on LB broth/plate

Teaching Experience

Fall 2014 Teaching assistant at Rutgers University

- Instructed General Biology 1 Lab. Ratings from students can be found at (download then open) https://drive.google.com/file/d/0B4LgXBHC0oEXMmdEOGp6OXFqRzdzamtDbWlDT3pYSXo4Z0tJ/view?usp=sharing

Presentations

1. Poster presentation at RECOMB/ISCB conference hosted by the International Society for Computational Biology: E-Flux2 and SPOT: Validated methods for inferring intracellular metabolic flux distributions from transcriptomic data, Philadelphia, PA, USA, Nov 15-18, 2015

2. Poster presentation at Metabolic Engineering X conference hosted by the International Metabolic Engineering Society and the Society for Biological Engineering - AIChE Technological Communities: Integration of Transcriptomic Data in Genome-scale Metabolic Models Predicts in Vitro Intracellular Central Carbon Metabolic Fluxes with High Correlation in Escherichia Coli and Saccharomyces Cerevisiae, Vancouver, BC, Canada, June 15-19, 2014

3. Poster presentation at 9th Annual Great Lakes Bioinformatics Conference hosted by the International Society for Computational Biology: Integration of Transcriptomic Data in Genome-scale Metabolic Models Predicts in Vitro Intracellular Central Carbon Metabolic Fluxes with High Correlation in Escherichia Coli and Saccharomyces Cerevisiae, Cincinnati, OH, USA, May 16-18, 2014

Awards and Scholarships

2017 Best Student Paper Award 2016, 1st prize, CCIB, Rutgers University, $250
2012-2017 Graduate/Teaching assistantship, Rutgers University, full tuition
2015-2016 TA/GA Professional Development Award, $700
2011 Outstanding Poster Award, The 2nd Asian Congress on Schizophrenia Research
2009-2010 Lecture and Research Scholarship, Seoul National University, full tuition
2008 Academic Excellence Scholarship, Korea University, full tuition

Professional membership

Member, International Society for Computational Biology (ISCB)

Professional development

- Took online courses offered by accredited universities to acquire knowledge and skills useful in enhancing research performance

2016 Genomics and Other Omics: The Comprehensive Essentials, **Stanford University**
(**link to certificate:**
https://drive.google.com/file/d/0B4LgXBHC0oEXZXBnYkZWSzl2RFU/view?usp=sharing)
2015 R Programming, **Johns Hopkins University**
(https://www.coursera.org/account/accomplishments/certificate/Q2MJM3VGB8)
2015 The Data Scientist's Toolbox, **Johns Hopkins University**
(https://www.coursera.org/account/accomplishments/certificate/Z7NMBYHE7K)
2014 Model Thinking, **University of Michigan**
(https://www.coursera.org/account/accomplishments/certificate/KG7ZCULWFR)
2014 Think Again: How to Reason and Argue, **Duke University**
(https://www.coursera.org/account/accomplishments/certificate/HWZRW72RU3)