# Flux Balance Analysis of *Escherichia coli* under Temperature and pH Stress Conditions

Thesis by

**Xiaopeng Xu**

In Partial Fulfillment of the Requirements

For the Degree of

**Masters of Science**

King Abdullah University of Science and Technology, Thuwal,

Kingdom of Saudi Arabia

May, 2015

The thesis of Xiaopeng Xu is approved by the examination committee

Committee Chairperson: Xin Gao

Committee Member: Victor Solovyev

Committee Member: Vladimir Bajic

# ABSTRACT


## Flux Balance Analysis of *Escherichia coli* under Temperature and pH Stress Condition

Xiaopeng Xu

An interesting discovery in biology is that most genes in an organism are dispensable. That means these genes have minor effects on survival of the organism in standard laboratory conditions. One explanation of this discovery is that some genes play important roles in specific conditions and are essential genes under those conditions. *E. coli* is a model organism, which is widely used. It can adapt to many stress conditions, including temperature, pH, osmotic, antibiotic, etc. Underlying mechanisms and associated genes of each stress condition responses are usually different. In our analysis, we combined protein abundance data and mutant conditional fitness data into *E. coli* constraint-based metabolic models to study conditionally essential metabolic genes under temperature and pH stress conditions. Flux Balance Analysis was employed as the modeling method to analysis these data. We discovered lists of metabolic genes, which are *E. coli* dispensable genes, but conditionally essential under some stress conditions. Among these conditionally essential genes, *atpA* in low pH stress and *nhaA* in high pH stress found experimental evidences from previous studies. Our study provides new conditionally essential gene candidates for biologists to explore stress condition mechanisms.

# ACKNOWLEDGEMENTS

I want to acknowledge my supervisor, Professor Xin Gao, for sponsoring my thesis studies at KAUST, and thank Dr. Hiroyuki Kuwahara, for his instructions throughout this project. My thesis would not have been possible without the help from both of them.

I would like to thank Professor Victor Solovyev, Professor Antoine Vigneron, and Professor Xiangliang Zhang. I benefited a lot about computational biology knowledge from the courses by Professor Victor Solovyev in my second semester at KAUST. The combinatorial optimization techniques and machine learning algorithms, which I learnt during my attendance of the excellent courses by Professor Antoine Vigneron and Professor Xiangliang Zhang, are very useful for me to conduct this project. So, I want to thank them for the valuable knowledge they taught me.

I also want to acknowldge my peer students and laboratory colleages for the valuable suggestions and advices during writing of my thesis.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| CBM | constraint-based metabolic model |
| FBA | flux balance analysis |
| FVA | flux variability analysis |
| LP | linear programming |
| MFV | mutant flux variance |

# LIST OF FIGURES

# LIST OF TABLES

# Nomenclature

**constraint-based metabolic model**

Constraint-based metabolic model is a mathematical model, which is transformed from a metabolic reconstructions, to model metabolism of an organism. In our study, the constraint-based models of *E. coli* mutants are converted from a *E. coli* metabolic reconstruction.

**fitness data**

Fitness describes the capacity of an organism to survive and grow under given environment. In our case, fitness data are normalized values from colony sizes of *E. coli* mutant strains under each condition.

**flux balance analysis**

Flux balance analysis is a mathematical method to study metabolism using genome-scale constraint-based metabolic model. It includes many programming methods, including standard flux balance analysis, which optimize flux of a single reaction, and flux variability analysis.

**flux variability analysis**

Flux variability analysis is a computational method in flux balance analysis, which uses standard flux balance analysis and calculate the minimum and maximum fluxes for all reactions in constraint-based metabolic model.

**linear programming**

Linear programming is a mathematical model which achieves the best outcome with requirements represented by linear relationships. Linear programming can be solved in nearly-linear time using approximate algorithms.

**metabolic network reconstruction**

A metabolic network reconstruction is a knowledgebase which contains all metabolic reactions of an organism and genes that express the enzymes. It can be readily transfromed into a constraint-based mathematical model to simulate a molecular mechanism of an organism.

**mutant**

Mutant is an organism or a strain, in which DNA sequence is changed or mutated. In this paper, a gene mutant is a *E. coli* strain with deletion or modification of the gene.

**mutant flux variance**

Mutant flux variance is a value defined for each mutant under specific stress condition group in our study. It is calculated as mean of flux variances of reactions over several conditions.

**protein abundance**

Protein abundance defines the relative abundance of each protein over the entire expressed proteome. Abundances of most proteins are believed to be intrinsic. That is, they are in a relatively stable range, which ensure functional effectiveness.

**stress condition**

Stress condition is an environment, where organisms need to response to environment demands or pressures to survive. Usually, organisms show a reduction in performance or fitness under stress conditions.

# Chapter 1

# Introduction

One of the most striking discoveries in molecular genetics is that most genes have minor effects on survival of an organism under standard laboratory conditions. That is, majority of genes in an organism are dispensable, while only a small portion of genes are essential [1]. This raises the question of why so many genes are dispensable.

To answer this question, people proposed that effects of gene deletion on phenotype and fitness depend on the environment of the organism, compensatory mechanisms for the gene, or both [1, 2]. Influence of environment on mutants, which is called phenotypic plasticity, is considered to be caused by dispensable genes [1, 2], while showing negligible effects in normal condition, may have significant contributions to fitness in a specific condition. Existance of compensation mechanisms is also called genetic robustness. Two compensation mechanisms are proposed for genetic robustness, existence of duplicated genes or isoenzymes, and compensation from alternative metabolic pathways and regulatory networks [1, 3]. Besides, phenotypic plasticity and genetic robustness are not mutually exclusive, they may work together to determine the phenotype of an organism [2].

Many mutation studies have been conducted in *E. coli* to learn essential genes and dispensable genes [4–6]. From these study, it is found that some genes may be essential under specific conditions [5, 7, 8] and dispensable in other conditions. *E. coli* has the ability to adapt to different stress conditions. These conditions include temperature,

pH, antibiotic, osmotic stress, etc. This raises an interesting question of conditionally essential and dispensable genes in these stress conditions for *E. coli*.

*E. coli* growing in temperature stress conditions and pH stress conditions displays some interesting effects, which are widely studied. For temperature conditions, balanced growth of *E. coli* can be achieved between 49℃ and 10℃ [9], however, the metabolic state and steady state level of proteins are quite different in different ranges [10, 11]. Within normal condition range (20℃ - 37℃), concentration levels of protein do not change significantly, suggesting that the metabolic state are very similar within this range. However, when temperature changes to high range (above 40℃) or low range (below 20℃), more profound cellular physiological changes occur, suggesting quite different metabolic states among these temperature ranges [11, 12]. For pH conditions, *E. coli* has the ability to adapt to a wide pH range without significant change of cytoplasmic pH [13–16]. The pH of *E. coli* changes only between 7.2 and 7.8, while external pH changes from 5.5 to 9, which indicates a mechanism which *E. coli* employs to maintain a pH homeostasis [13].

Most of previous computational investigations of *E. coli* under temperature stress conditions focus on mathematical modeling of regulation during heat shock response [17–20]. In 2002, a metabolic flux study was conducted by Jan Weber et al. [21] to learn the flux redirection of recombinant *E. coli* upon temperature upshift. None has compared the metabolic fluxes under different temperature ranges. Similarly for pH stress condition, most *E. coli* pH stress studies focus on pH homeostasis mechanisms in *E. coli*, without considering holistic picture of influence of pH stress on metabolism [14, 15].

In 2011, Nichols et al. published a comprehensive experimental phenotypic study of *E. coli* mutants under stress conditions, including temperature, pH, osmotic, and antibiotic/antimicrobial stress conditions[8]. In their study, they determined fitness values for the Keio gene deletion library of *E. coli K-12 MG25113* [22]; essential

gene hypomorphs [23]; and a small RNA/protein knockout library [24] in 324 stress conditions. Among the 324 conditions, 7 are temperature conditions and 8 are pH conditions. *E. coli* mutant strains are growing under 4 high temperature conditions (40℃, 42℃, 43.5℃, and 45℃), 3 low temperature conditions (16℃, 18℃, and 20℃), 4 high pH conditions (8, 9, 9.5, and 10), and 4 low pH conditions (4, 4.5, 5, and 6) to determine their colony sizes under these conditions. The colony sizes were observed and converted into mutant fitness values, which are comprised of fitness values for the 3979 mutants.

In this study, we combined both *E. coli* protein abundance data [25] and the mutant fitness data to study *E. coli* metabolic fluxes under cold conditions, heat conditions, high pH conditions, and low pH conditions. The *E. coli* strain we analyzed is *E. coli K-12 MG25113*. A genome-scale *E. coli* metabolic network reconstruction metabolic reconstruction, iJO1366 [26], was employed to analyze the flux changes of *E. coli* mutants under different temperature and pH conditions. A constraint-based modeling method, flux balance analysis[27], was used in our mathematical modeling.

We defined conditionally essential and dispensable genes based on experimental mutant fitness values under certain stress condition. A variable, mutant flux variance (MFV), was used to represent the flux variances of a mutant over certain stress conditions. A significant correlation between mutant flux variance (MFV) and conditional gene essentiality and dispensability. Conditionally essential genes tends to have small MFVs, while conditional dispensable genes have large or small MFVs. Functions of the small MFV conditionally essential genes are checked in literature, and some experimental evidence of conditional essentiality is found.

## 1.1 Objectives and Contributions

The main objective of our study is to use a constraint-based metabolic modeling method, flux balance analysis, to analyze *E. coli* metabolic fluxes under four different stress condition groups, cold conditions, heat conditions, low pH conditions, and high pH conditions. 1249 non-essential gene mutants are studied in our process.

The contributions of this thesis folds in the following streams:

• Constructed mutant constraint-based metabolic models from metabolic reconstruction of *E. coli*. Protein abundances of proten products are considered while building mutant constraint-based models, which refine traditional knockout effect of mutated genes.

• Converted mutant fitness data under different stress conditions into growth rate data, which are used in flux variability analysis.

• Conducted flux variability analysis for each mutant under each condition to get flux distributions on reactions.

• Analyzed the flux distributions of mutants under different conditions, and constructed a variable, mutant flux variance, to represent the flux variances of a mutant under certain kind of stress.

• Analyzed mutant fitness data and defined sets of conditionally essential and dispensable genes for each kind of stress.

• Discovered a list of conditionally essential metabolic genes, which show significant low mutant flux variances, in each condition group. These conditionally essential genes can be potential targets for biologists to study metabolic mechanisms under temperature and pH stress conditions.

# Chapter 2

# Background

## 2.1 Metabolic Network and Modeling

### 2.1.1 Metabolic Network

Cell has the ability to break down some molecules and synthesize some other molecules. This biological process consists of a complex but highly regulated network of chemical reactions, which are collectively called metabolism. These chemical reactions, which are usually catalyzed by enzymes, are also called metabolic reactions. They allow a cell to maitain its structure, grow and replicate, and respond to its environment.

The chemical reactions in a cell are connected into several metabolic pathways, in which some chemicals are consumed, while other chemicals are produced along the series of chemical reactions. These chemicals are called metabolites, which are the intermediates and products of metabolism. In a metabolic pathway, a product of one reaction involves the next reaction as a substrate. Also, the chemical reactions of a metabolic pathway are catalyzed by a sequences of enzymes.

Enzymes are macromolecular biological catalysts, which catalyze chemical reactions. In order to take place at a fast rate, almost all chemical reactions in the cell require enzymes. But a chemical reaction can be catalyzed by multiple enzymes and one enzyme can involved in different chemical reactions. The set of enzymes existing

in a cell determines what chemical reactions and which metabolic pathways appear in that cell. Through changing the set of available enzymes, a cell can the regulate the set of metabolic pathways in response to environment changes of the cell or to chemical signals from the neighboring cells.

Almost all enzymes of metabolic reactions are proteins. Following the central dogma of molecular biology, that is, DNAs are transcribed into RNAs and RNAs are translated into proteins, the proteins in a cell are determined by DNAs. A gene is a hereditary unit composed of a sequence of DNA. One enzyme may controlled by one gene or multiple genes. The availability of an enzyme is determined by availability and expression of its corresponding genes. Protein product of one gene may involve in multiple enzymes. Knockout of a gene causes the decreased-concentration or missing of related enzymes in a cell.

A metabolic network is the complete set of metabolic reactions and physical processes in a cell. It contains all chemical reactions of a cell, regulatory interactions that control those reactions, and genes that are map into reactions. From a metabolic network, we can learn the metabolism and physiological properties of a cell.

### 2.1.2 Metabolic Network Modeling

Traditional modeling of multiple enzymatic reactions usually uses enzyme kinetic models. These models include mass action kinetics [28], Michaelis-Menten kinetics [29], Hill equation [30], ternary-complex mechanisms [31], and ping-pong mechanisms [32]. Mass action kinetics is one of the most common kinetic assumptions in enzyme kinetics. The rate of a chemical reaction is assumed to be proportional to the product of reacting chemical species concentrations. Mass action assumption is used in many kinetic models, including Michaelis-Menten kinetics and Hill equation. Single substrate enzymatic reactions are often modelled using Michaelis-Menten kinetics, which can also be applied into multi-substrate modeling with some conditions. Hill equation

is used to model macromolecule and ligand binding. Ternary-complex mechanisms and ping-pong mechanisms are used to describe multi-substrate enzymatic reaction. However, in order to use these enzyme kinetic models, kinetic parameters of reactions need to be supplied.

A metabolic network contains all metabolic reactions of a cell. Traditional kinetics-based metabolic modeling methods model a series of metabolic reactions using coupled differential equations. Enzyme kinetic parameters and metabolite concentrations are required in thoses modeling methods. As a lack of kinetic parameters and metabolite concentration information, it is impossible to model all chemical reactions of a cell using those methods. In order to model the metabolic network of a cell, constraint-based metabolic network modeling methods are proposed. These constraint-based metabolic modeling methods include flux balance analysis [27], extreme pathways [33], elementry mode analysis [34], etc. Flux balance analysis is a constraint-based modeling methods which is most widely used [35].

## 2.2 Constraint-Based Metabolic Modeling of *E. coli*

### 2.2.1 Metabolic Network Reconstruction

In systems biology, in order to systematically study the metabolism of an organism, genome-scale metabolic network reconstructions are widely used. A network reconstruction is a structured knowledgebase that contains detailed biochemical, genetical, and genomical information about an organism [36]. It has become a denominator in systematical study of an organism [26]. Metabolic network reconstruction contains detailed metabolic network information, including chemical formulas and charges of metabolites, stoichiometric parameters of metabolic reactions, and gene, protein, and reaction associations. It is an indispensable tool to study metabolism of an organism in system-level. Many metabolic network reconstructions have been constructed in

the last several years [27], which contain all metabolic reactions of an organism and enzyme-coding genes associated with reactions.

### 2.2.2 Flux Balance Analysis

Flux balance analysis (FBA) is a widely used mathematical method for studying metabolism using genome-scale metabolic reconstructions. It is able to simulate growth rate of an organism and predict the production of some important metabolites [27]. With metabolic reconsructions of many organisms constructed and high-throughput technologies speeding construction of metabolic reconstructions, FBA becomes an important tool for metabolism studies[37].

FBA is a constraint-based modeling method with two assmuptions made, steady state of metabolites in the cell, and targeted optimization of the organism. The steady state of metabolites assumption is that, metabolic system is in a steady state, in which the concentrations of metabolites do not change. This is because comparing to transcription and translation, metabolism is a much faster process. In a relatively small time-scale, we can consider metabolism to be in a pseudo-steady state. In this steady state, the production of a metabolite cancels out with the consumption of the metabolite. That is, the summation of fluxes producing a metabolite equals to the summation of fluxes consuming the same metbolite. In targeted optimization assumption, organism is assumed to have optimized some biological goals during evolution. Usually, the goal is defined as optimizing growth. A pseudo-reaction, biomass reaction, is added to the metabolic reconstruction, maximizing flux of which are usually used as the objective function during optimization.

A metabolic reconstruction includes a stoichiometric matrix and a set of constraints. Assume that $B$ is the set of metabolites and $R$ is the set of reactions, including biomass reaction, in a metabolic network. Dimension of stoichiometric matrix, denoted by $\mathbf{S}$, is $|B| \times |R|$, where $|B|$ and $|R|$ are the number of metabolites and

reactions in metabolic network respectively. A value $\mathbf{S}_{br}$ in the stoichiometry matrix denotes the stoichiometric coefficient of a metabolite $b$ in reaction $r$. If a metabolite is consumed in a reaction, the stoichiometric coefficient is set to be negative, otherwise, when a metabolite is produced in the reaction, the stoichiometric coefficient is set as positive. Assume that $\mathbf{V}$ is the vector of fluxes passing through reactions with dimension $|R| \times 1$. Based on steady state assumption, for each metabolite we have an equation, in which weighted summation of fluxes equal to zero. The weights are the stoichiometric coefficients, which describe the consumption and production of the metabolite in reaction. Then the system of equations is denoted as the dot product of stoichiometry matrix $\mathbf{S}$ and flux vector $\mathbf{V}$ equals to a zero vector of $|B| \times 1$ dimension.

Calculating the steady state flux distribution of metabolic reactions is formulated as a linear programming (LP). The mathematical formulation of this LP is

$$\begin{aligned} &\text{Maximize} && \mathbf{C}^{\mathrm{T}}\mathbf{V} \\ &\text{Subject to} && \mathbf{S} \cdot \mathbf{V} = \mathbf{0} \\ & && l_r \leq v_r \leq u_r, \forall r \in R. \end{aligned}$$

Where, $\mathbf{C}$ is a known vector of contants corresponding to weight of each fluxes in linear programming objective function. $r$ is a reaction in the metabolic network. $v_r$ represents flux of reaction $r$. $l_r$ and $u_r$ are the lower bounds and upper bounds of flux passing through reaction $r$.

Other constraints are added into FBA in the form of flux bounds of reactions, namely $l_r$ and $u_r$ in previous equation. $l_r$ of all irreversible reactions are set to be 0 and $l_r$ of reversible reactions can be negative values. Maximum value of $u_r$ is 1000, and minimum value of $l_r$ is -1000. Lower bounds of exchange reactions are set as zero for the metabolites that are not present in the growth medium. Lower flux bounds of exchange reactions, whose metabolites are supplied in the growth medium, are set

to be negative values. Besides, lower and upper flux bounds of the some reactions are constrained to meet experimental conditions, such as low enzyme activites and knockouts of metabolic genes.

Comparing to traditional metabolic modeling methods, FBA needs less input data to construct the model and are more computationally efficient as it uses LP as its optimizing method [38]. Also the growth simulation from FBA fits very well with biological experiments. In biological study, FBA has wide applications. It can be used in bioprocess engineering to recognize important metaboic reactions in metabolic reconstructions of microbes, which are used for fermentation and production of important chemicals, such as ethanol and drug precursors, in a systematic way [39]. Also, it is be used to discover putative drug targets in some complex diseases, such as cancers [40]. Besides, FBA is used to simulate some complex metabolic and evolutionary effects, such as phenotypic plasticity in yeast [1], Warburg effect in cancer [41], and host-pathogen interaction [42].

### 2.2.3 Omics Data and Genome-Wide Metabolic Reconstructions

During the past decade, high-throughput technologies produced a massive amount of biological data. Genomics, transcriptomics, and proteomics enable researchers to get a lot of data about many biological phenomena, including obesity, cancer, infection, biofuel production, and host-pathogen interaction [43]. Many statistical inference methods are applied to study these omics data to learn the important genes and proteins for these biological phenomena. However, these methods usually do not explain the underlying mechanisms. Increasing number of biological knowledgebases are developed to extract the biological knowledge and integrate these omic data [44–46].

High-throughput technologies provide researcher with measurement of large num-

bers of molecules, such as proteins, metabolites, DNA and RNA, These molecules interact in a network of interactions to produce cellular functions and determine the phenotypes of an organism. However, to extract knowledge from the ocean of omics data is not trivial. Deficiency of data, difference between experimental platforms, and few experimental validation for hypotheses testing result in a serious lagging of analysis efforts [47–50].

In order to tackle the challenge of big omics data, knowledge-based methods are widely employed. Knowledge-based methods are based on the annotation of omics data, and construct networks from biochemical and genomic data. Examples of such methods include KEGG [44], EcoCyc [45], and constraint-based metabolic reconstruction [46]. Network reconstructions in KEGG and EcoCyc contain useful information, however, cannot be easily transformed into a mathematical model to be used in biology study. Constraint-based reconstructions not only contain information existed in the networks, but also can readily be transformed into mathematical models, thus provide a good way for biological study of biological networks.

Network reconstructions combine existing biochemical, genetic, and omics data to generate a knowledge network of an organisms molecular components with their interactions as edges [51, 52] Though network reconstructions can be employed to construct regulatory [53] and signaling [54] networks, the greatest success was in reconstruction of metabolic networks. Difficulties in genome-scale modeling of regulation and signaling make application of signaling and regulation reconstruction more limited than metabolic network reconstruction. Decades of biochemical research leaves a huge legacy in metabolic network research and a simple mathematical model, flux balance analysis, makes it very efficient to study genome-scale metabolic reconstructions.

There are two general approaches to utilize omics data in metabolic network reconstruction [43, 46]. Omics data can be utilized to reconstruct a metabolic network model, or integrated into an existing metabolic network reconstruction; the predic-

tion from metabolic network reconstruction analysis can be compared and validated using omics data. For example, genomic data can be employed to infer enzyme encoding genes in an organism. Based on these enzyme encoding genes, corresponding reaction network can be constructed. Transcriptome data can be used as constraints in constraint-based metabolic models to create condition- or tissue-specific model, which can be utlized to study condition- and tissue-specific mechanisms.

### 2.2.4 History of *E. coli* Metabolic Reconstructions

*E. coli* is extensively studied in metabolic reconstruction and has the most comprehensive and complete metabolic reconstruction than any other organisms [26]. *E. coli* metabolic reconstruction is based on *Escherichia coli K-12 MG1655* . The first metabolic reconstruction for *E. coli* is iJE660, which was published in 2000 [55]. Information are gathered through extensively searching in literature and database to reconstruct this network. A later released *E. coli* construction, iJR904, was published in 2003 [56]. Alternative carbon source pathways and specific quinone utilization pathways were included in this model. Also gene-reaction associations relationship was incorporated in iJR904, with many new genes and reactions added. The *E. coli* network was expanded again in the next update, iAF1260. In iAF1260, many reactions and metabolites were located to their corresponding cellular location [57]. Besides, in iAF1260, the lower bounds on irreversible reactions were modified based on the their thermodynamic properties. In 2011, the latest *E. coli* metabolic network reconstruction, iJO1366 was published, which includes 1136 unique metabolites, 2251 metabolic reactions, and 1366 metabolic genes [26]. In this update, several genes were added and gaps in previous reconstruction were identified and removed.

In our analysis, we used iJO1366 *E. coli* reconstruction, because it is the most complete and comprehensive *E. coli* reconstruction and during the construction of iJO1366, metabolism in other *E. coli* species are also considered. Thus iJO1366

can be easily adapted to construct metabolic network reconstruction of other *E. coli* strains, which is required in our analysis.

## 2.3 *E. coli* Stress Condition Effects

### 2.3.1 *E. coli* Responses under Temperature Stress Conditions

For cold temperature conditions, the most studied *E. coli* effect is the cold shock response. Cold shock response happens when the temperature condition of *E. coli* is downshift to a low temperature, for example from 37℃ to 10℃ [9]. During acclimation phase of cold shock response, several cold shock proteins (CSPs) are overproduced temporarily to help *E. coli* adapt to low temperature [58]. After the acclimation phase is over, the CSPs decrease to a new basal level and expression of other proteins resumes, which allows *E. coli* to grow at a slower rate under low temperature [11, 59].

The most extensively studied effect of *E. coli* growing under high temperature condition is heat shock response. Heat shock response comes when *E. coli* is treated with a sudden upshift of temperature. A set of highly conserved proteins – heat shock proteins (HSPs) and an alternative sigma factor – σ32 are considered main characters in heat shock response [60]. The heat shock response can be classified into three phases, which are induction phase, adaptation phase, and steady state phase [61]. During induction phase, increased translation, decreased negative regulation, and transient stabilization of σ32 result in a rapidly increased σ32 abundance and activity, thus more HSPs are induced. Increased σ32 level and activity also induce more protease and molecular chaperones, which negatively regulates σ32 level [62]. In adaptation phase, this feedback regulation regulates the cell to go to a relatively steady state.

### 2.3.2 *E. coli* Mechanisms under pH Stress Conditions

*E. coli* has evolved several different protective mechanisms, which enable them to maintain a nearly steady cytoplasic pH level in otherwise life-threatening pH environments. In low pH conditions, two types of adaptations defined for bacteria, ATR and XAR [63, 64]. ATR indicates a adaptation in mildly acidic pH that enhances survival in severe acidic environments. XTR allows unadapted cells to survive in pH which are too acidic (pH $\leq$ 2.5) to permit growth. *E. coli* employs both two types of adaptations. $F_1F_0$-ATPase machine, amino acid-dependent decarboxylase/antiport systems, deiminase and deaminase systems, repair and damage prevention of proteins, and modification of cell memberane enable *E. coli* to adapt to acidic environments[14, 65].

In high pH conditions, *E. coli* has four different strategies to maintain a pH homeostasis. These strategies include increasing metabolic acid production by sugar fermentation and amino acide deaminases, increasing ATP synthase activity which couples with $H^+$ importing into cytoplasm, increasing synthesis and activity of monovalent cation/proton antiporters, and cell surface property change [15, 16]. Among the four strategies, mechanism of monovalent cation/proton antiporters plays a major in *E. coli*'s adaptation to alkaline pH stress. Specifically, $Na^+/H^+$ antiporter *nhaA* and multidrug transporter *mdfA* have dominate roles in alkaline pH homeostasis in *E. coli* [15].

### 2.3.3 Stress Conditions and Gene Essentiality

In biology theory, phenotypic plasticity is one of the two explanations for the striking discovery that most genes in an organism are non-essential [1, 2]. Phenotypic plasticity describes the ability of an organism, while holding a same genotype, can have multiple phenotypes when exposed to different environments [2, 66–68]. For a single gene, phenotypic plasticity means that this gene might be essential under cer-

tain environments, but dispensable in other conditions [4, 5, 8]. Phenotypic plasticity was widely studied in multi-cellular organisms [66–68]. To elucidate the mechanisms of phenotypic plasticity using simple model organisms, several studied in *E. coli* are also conducted [69–71]. In our study, we analyzed the conditionally essential genes in *E. coli* stress conditions. These genes might help us understand some phenotypic plasticity phenomena of *E. coli*, when faced with different stress conditions.

# Chapter 3

# Methods

## 3.1 Method Overview

We combined *E. coli* metabolic reconstruction, protein abundance data, and experimental mutant fitness data to analyze *E. coli* metabolic fluxes in stress conditions.

Protein abundance information was combined with *E. coli* metabolic reconstruction to construct the constraint-based metabolic models (CBMs) for mutants of metabolic genes. Traditional knockout of a gene was simulated as setting constraints of all reactions related the gene as zero. In our simulation, the flux constaints of reactions, which related to the knockdown or knockout gene, were changed based on protein abundance of genes related to that reactions. With the CBMs of mutants, FBA was run for each mutant with maximizing flux of biomass reaction set as objective function. Here, we got the optimal growth rate for each mutant, as shown in Figure 3.1 A.

Flux variability analysis (FVA) was conducted for each mutant under each temperature and pH stress conditions. First, mutant fitness data were transformed into growth rates. The transformation was done using the optimal growth rates obtained for each mutants. Then, the growth rate was added into CBM of the mutant as constraints to study the flux distribution of a mutant under a certain condition. The constraints of biomass flux were modified with both lower and upper flux bounds

set to be the growth rate of the mutant under that condition. Then FVA was conducted for the modified model. The results are the minimum and maximum fluxes for reactions in the *E. coli* constraint-based model, as shown in Figure 3.2 B.

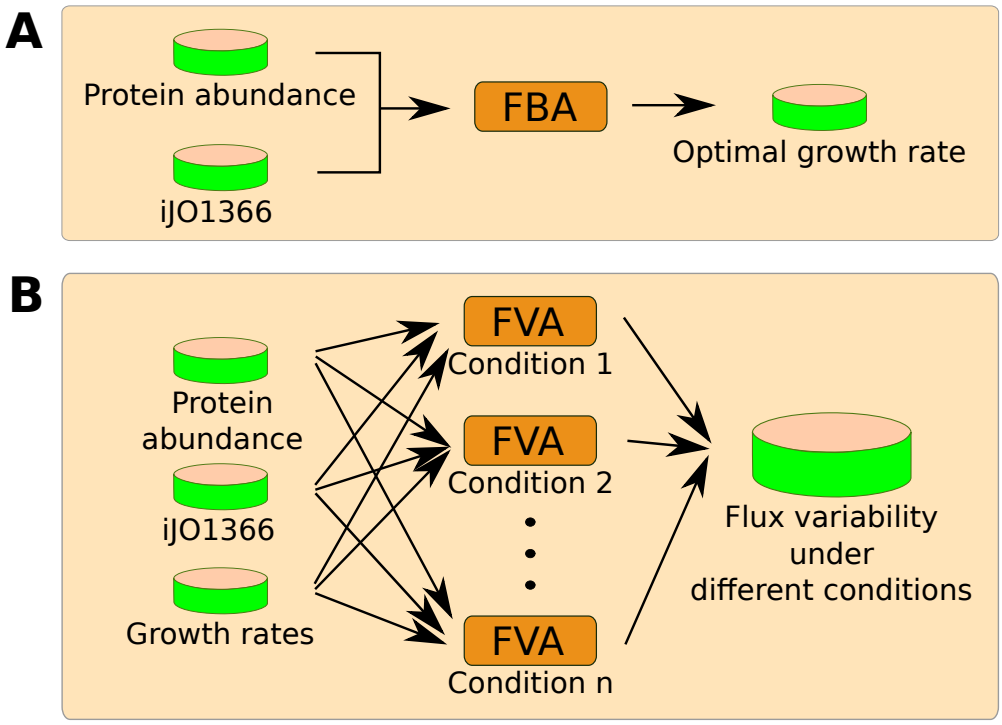


Figure 3.1: Overview of analysis. First, for a mutant, protein abundance data and *E. coli* iJO1366 metabolic reconstruction are used to conduct flux balance analysis. The two data are used to construct the mutant CBM. Flux balance analysis is conducted for this mutant CBM to get the optimal growth rate of this mutant. After that, mutant fitness data are transformed into growth rates using the optimal growth rate data. Finally, protein abundance data, iJO1366 metabolic reconstruction, and growth rate data are used to conduct flux variability analysis for a mutant under each condition. Growth rate data are added as constraints to conduct FVA.

With flux distribution calculated, we conducted a series of analyses to study fluxes distributions of mutants. To analyze the patterns in flux distribution, we extracted a value MFV. For each mutant, variances of maximum fluxes were calculated for each reactions in stress conditions, and average of the flux variances in all reactions was calculated as MFV. MFVs are used to compare mutants in different environmental

stress. Conditionally essential genes and dispensable genes were defined based on the fitness data. Genes with fitness values of corresponding mutant smaller than -1 in all conditions of a stress condition group, were defined conditionally essential genes of this stress. Genes, whose fitness values are larger than 1 in all conditions of a stress condition group, were defined conditionally dispensable genes of this stress. We found clear MFV distribution patterns of conditionally essential genes and conditionally dispensable genes. Then we checked the functions of the conditonally essential genes with low MFVs. Some experimental evidences were found for the genes we discovered.

For FBA of mutant CBMs, linear programming was conducted using GNU Linear Programming Kit [72]. FVA was carried out for all mutants under each stress condition using Cobrapy [73] to obtain fluxes of each reactions with certain growth rate.

## 3.2 Constraint-Based Model Construction for *E. coli* Mutants

To use constraint-based metabolic modeling methods in our analysis, several modifications of original *E. coli* metabolic reconstruction were conducted. The *E. coli* constraint-based metabolic models constructed in our study are for *E. coli* mutants from the study by Nichols [8]. The constraint-based metabolic models of mutants were based on the *E. coli* metabolic reconstruction – iJO1366, which is the most comprehensive *E. coli* metabolic network reconstruction [26].

### 3.2.1 Map Genes in Phenotype Data into *E. coli* Reconstruction

Model iJO1366 represents a metabolic network reconstruction of *E. coli K-12 MG1655*. Mutant fitness data is experimental data for *E. coli K-12 MG25113*. In order to model *E. coli K-12 MG25113*, we followed a work done by Orth et al. [74]. Seven reactions were removed from iJO1366 to model *E. coli K-12 MG25113*. To match with experimental medium, the default medium settings are modified to fit LB medium [75].

Then mutant genes in mutant fitness data were mapped to genes in iJO1366 model. The gene ID used in phenotypen data is ECK number, which is based on genome sequences of two *E. coli K-12* bacteria [76]. The gene ID used in iJO1366 model is B number, which is based on genome sequences of *E. coli K-12 MG1655* [77]. A mapping data [76] were used to map ECK numbers to B numbers. During this mapping, all genes in iJO1366 can find unique-matching ECK numbers. Through comparing the ECK numbers, 1249 out of 1366 metabolic genes in iJO1366 model have corresponding mutant phenotype data. Six of these genes have multiple mutants corresponding to them (i.e. ECK0086(3), ECK0097(3), ECK0179(2), ECK0905(2), ECK2019(2), and ECK3623(3)). 1258 out of 3979 mutants in mutant fitness data were mapped to metabolic genes in iJO1366 model.

### 3.2.2 Construct Constraint-Based Metabolic Model for Each Mutant

Most previous computational studies of a gene knockout using FBA simulate it by constraining the associated reactions to carry no fluxes [1, 27, 78, 79]. This simulation has a problem when expression of a gene is relatively low comparing to compensatory genes in its associated reaction, in which case, knockout of this gene has minor effects on the reactions. To quantatively estimate the effect of a gene knockout on

the metabolic fluxes, we used a comprehensive protein abundance data from PaxDb [25]. Protein abundance data are the relative abundance of gene expression products. Because most proteins are discovered to have relative stable abundance levels to be functionally effective, the protein abundance can be considered to be stable in different environmental conditions [25]. The protein abundance data we used are a weighted average data from 6 independent works [80–85]. The data contain the protein averaged abundance information of 3165 *E. coli* genes. 1151 genes in *E. coli* iJO1366 model have protein abundance information, 1043 of which have corresponding phenotype data. For other metabolic genes in iJO1366, which do not have the protein abundance information, median of all protein abundance values are used [41].

To evaluate the effect of gene knockout on reactions, enzyme activity was assumed to be in proportional to corresponding protein abundance [41]. For a metabolic reaction $r$, suppose it is associated with four genes $A$, $B$, $C$, and $D$. Protein products of $A$ and $B$ combine together as an enzyme $AB$ to catalyze reaction $r$. Protein products of $C$ and $D$ combine together to catalyze reaction $r$ as an enzyme $CD$. Here, $AB$ and $CD$ are two enzymes with same function, both of them can catalyze reaction $r$ independently. The rate of flux constraint change after gene $A$ knockout is calculated to be summation of abundance of $C$ and $D$ over summation of abundances of $A$, $B$, $C$, and $D$, as shown below.

$$l_r(\neg A) = \frac{c+d}{a+b+c+d} \cdot l_r$$
$$u_r(\neg A) = \frac{c+d}{a+b+c+d} \cdot u_r$$

where, $a$, $b$, $c$, and $d$ are protein abundances of four genes $A$, $B$, $C$, and $D$ respectively. $l_r$ and $u_r$ are the lower and upper flux bounds of reaction $r$.

For example, metabolic reaction *ARGabcpp* is associated with nine genes *argT*, *hisM*, *hisP*, *hisQ*, *artI*, *artJ*, *artM*, *artP*, and *artQ*. Protein products of *argT*, *hisM*,

*hisP*, and *hisQ* combine together as an enzyme complex I to catalyze this reaction. Protein products of *artI*, *artJ*, *artM*, *artP*, and *artQ* combine together as an enzyme complex II to catalyze this reaction. Here,these two enzyme complexes have the same function, both of them can catalyze reaction *ARGabcpp* independently. The rate of flux constraint change after gene *argT* knockout is calculated to be summation of protein abundance of *artI*, *artJ*, *artM*, *artP*, and *artQ* over summation of abundances of these nine gene products, as shown below.

$$l_{ARGabcpp}(\neg argT)=\frac{arti+artj+artm+artp+artq}{argt+hism+hisp+hisq+arti+artj+artm+artp+artq}\cdot l_{ARGabcpp}$$

$$u_{ARGabcpp}(\neg argT)=\frac{arti+artj+artm+artp+artq}{argt+hism+hisp+hisq+arti+artj+artm+artp+artq}\cdot u_{ARGabcpp}$$

where, $argt$, $hism$, $hisp$, $hisq$, $arti$, $artj$, $artm$, $artp$, and $artq$ are protein abundances of gene *argT*, *hisM*, *hisP*, *hisQ*, *artI*, *artJ*, *artM*, *artP*, and *artQ* respectively. $l_{ARGabcpp}$ and $u_{ARGabcpp}$ are the lower and upper flux bounds of reaction *ARGabcpp*.

The *E. coli* phenotype data produced by Nichols [8] contain 3979 mutant strains. 1258 mutants corresponds to metabolic genes in iJO1366 model. For each of 1258 metabolic gene-related mutants, a constraint-based metabolic mutant models was constructed. The construction of mutant models was based on *E. coli* metabolic model – iJO1366. In order to suit the model to a mutant and construct mutant FBA model, lower and upper flux bounds, that is $l_r$ and $u_r$, of reactions related the mutant were changed. Assume, $M$ was the set of mutants related to metabolic genes. For a mutant $m \in M$, the corresponding metabolic gene in *E. coli* model was found (as described in section 3.2.1) . Then associated reactions of the gene is retrieved based on iJO1366 model. If the gene is associated with multiple reactions, constraints of all these reactions were changed based on the protein abundence as described above. Then we got mutant CBMs for all metabolic gene mutants.

## 3.3 Flux Balance Analysis of *E. coli* Mutants

### 3.3.1 Conduct Flux Balance Analysis for *E. coli* Mutants

The formulation of FBA to optimize flux of biomass reaction, which is called optimal growth rate, is as follows.

$$\begin{aligned} &\text{Maximize} && v_{growth} \\ &\text{Subject to} && \mathbf{S} \cdot \mathbf{V} = \mathbf{0} \\ & && l_r \leq v_r \leq u_r, \forall r \in R. \end{aligned}$$

As mentioned before, $\mathbf{S}$ is the stoichiometry matrix of all reactions in *E. coli* model. $\mathbf{V}$ is an vector of fluxes of all reactions. $\mathbf{0}$ is an vector of 0, which means the metabolites are in steady state. $R$ is a set of all reactions in the model. $v_r$ is the flux of a reaction $r$ in $\mathbf{V}$. $v_{growth}$ is the flux of biomass reaction, which is a flux value in $\mathbf{V}$. Elements of $V$, $v_r$, are the variables. $\mathbf{S}$, $R$, and original $l_r$ and $u_r$ for each reaction $r \in R$ is already given in iJO1366 model.

Mutant specific CBMs were used with maximizing flux of biomass reaction set as objective function. FBA was run for each mutant $m \in M$ to get maximum flux of biomass reaction. The result is the maximum flux of biomass reaction, which is called optimal growth rate, for a mutant. In later notation, the optimal growth rate for a mutant $m$ in $M$ was denoted as $G_m$.

### 3.3.2 Convert Mutant Fitness Data into Growth Rate Data

The *E. coli* phenotype data produced by Nichols et al. [8] contain mutant fitness values of 3979 mutant strains, most of which are from Keio collection [22], under 324 stress conditions. In order to incorporate mutant fitness data in models, a transformation from mutant fitness values into growth rates was conducted. In mutant fitness

data, the mutant fitness values were calculated by converting colony sizes in certain condition into a standard normal distribution. As colony sizes are in proportional to the growth rates with the assumption that colony sizes are cumulated growth rates [41], mutant fitness values, which are normalized colony sizes, also are in proportional to the growth rates in each condition. Mutant fitness values were linearly transformed into growth rates to maintain the information contained in this condition.

During this mapping, two parameters are required for each condition. As shown below, $A_c$ and $B_c$ are two parameters required for condition $c$. To conduct this mapping, some constraints need to be fulfill. For a mutant $m$ under certain condition $c$, the mapped growth rate $g_{mc}$ must be no smaller than 0 and no greater than optimal growth rate of the mutant $G_m$. In order to determine these two parameters, two mapping relationships are used: the smallest mutant fitness value under each condition was mapped to 0 growth rate; the largest mutant fitness value under each condition was mapped to the optimal growth rate of a mutant, which has the largest mutant fitness value under this condition.

$$A_c * \max_m\{f_{mc}\} + B_c = G_{m^0}$$

$$A_c * \min_m\{f_{mc}\} + B_c = 0.$$

Where, $C$ is the set of all stress conditions. $c$ is a condition in $C$, $M$ is the set of all mutants of metbolic genes. $m$ is a mutant in $M$. $f_{mc}$ is the fitness value for mutant $m$ under stress condition $c$. $m^0$ is the mutant, which has the maximum fitness value $f_{mc}$ under stress condition $c$. $G_m$ is the optimal growth rate for a mutant $m$ and is calculated from FBA.

For a mutant $m$ under a condition $c$, growth rate was calculated based on fitness value using these two parameters $A_c$ and $B_c$. Assume $g_{mc}$ is the transformed growth rate for mutant $m$ under conditions $c$. To conduct FVA, the $g_{mc}$ must be smaller than

or equal to optimal growth rate of the mutant, $G_m$, otherwise, no flux distribution can achieve this growth rate.

$$0 \leq g_{mc} \leq G_m, \forall c \in C, \forall m \in M.$$

In order to ensure that $g_{mc}$ is achievable, another constraint was added. The growth rate $g_{mc}$ was calculated as the minimum of calculated growth rate from the mapping above and $G_m$. By calculating growth rate for each mutant under each condition based on fitness values, we got a dataset of growth rates for all mutants under these stress conditions.

$$g_{mc} = \min\{A_c \cdot f_{mc} + B, G_m\}$$

### 3.3.3 Conduct Flux Variability Analysis for Mutants under Stress Conditions

FVA is a computational method included in FBA. It uses standard FBA and calculate lower and upper bounds of fluxes for all reactions in a CBM. For a mutant $m$ under certain condition $c$, we conducted the FVA using mutant models. $\mathbf{S}$, $R$, and $l_r$ and $u_r$ for any reaction $r$ in $R$ are already known. Mutant specific constraints for biomass were added by setting lower and upper bounds of biomass flux $v_{growth}$ to growth rate $g_{mc}$.

During FVA, for each reacion $k$ in model, two linear programming are conducted,

as shown below.

$$\begin{aligned} &\text{Maximize/Minimize} && v_k \\ &\text{Subject to} && \mathbf{S} \cdot \mathbf{V} = \mathbf{0} \\ & && l_r \leq v_r \leq u_r, \forall r \in R \\ & && v_{growth} = g_{mc} \end{aligned}$$

$2|R|$ LPs were conducted during FVA for a mutant $m$ under condition $c$, where $|R|$ is the number of reactions. We conducted LPs for all $|M|$ mutants under $|C|$ stress conditions. FVA was conducted using Cobrapy package [73]. The data we get from FVA are minimum and maximum fluxes of each reaction in the model for $|M|$ mutants under $|C|$ conditions.

During FVA under stress conditions, we chose specific conditions, temperature and pH stress conditions, to study. Temperature and pH stress conditions were classified into 4 stress condition groups based on underlying stress conditon response. These stress condition groups include cold stress conditions, heat stress condition, low pH stress condition, and high pH stress condition.

To compare flux distribution differences of mutant growing under these stress conditions, we constructed a variable, MFV. For each mutant, variances of maximum fluxes were calculated for each reactions in different conditions of a stress condition group. For example, for cold stress condition group, flux variances of each reaction of a given mutant were calculated over three conditions, 16℃, 18℃, and 20℃. Then, for a mutant, average of all flux variances wss calculated as MFV to compare mutants under certain stress conditon group.

## 3.4 Gene Essentiality Study in Stress Conditions

### 3.4.1 Define Conditionally Essential and Dispensable Genes

In order to check the correlation between gene essentiality and MFVs under each stress conditon group, we conducted a statistical analysis of the genes. We defined conditionally essential genes and conditionally dispensable genes from *E. coli* dispensable genes based on fitness values. Dispensable genes are the genes, which are dispensable in standard laborotary conditions. Dispensable genes with fitness scores of corresponding mutant smaller than -1 in all conditions of a stress condition group, were defined conditionally essential genes of this stress. Dispensable genes, whose fitness scores are larger than 1 in all conditions of a stress condition group, were defined conditionally dispensable genes of this stress.

For each condition group, metabolic genes were divided into 10 groups based on MFVs. Each group contain 10 % of total number of mutated metabolic genes. G1 contains genes with MFVs in smallest 10%. G2 contains genes with MFVs between 10% to 20%. G10 contains genes with MFVs in largest 10%. Numbers of conditionally essential genes and conditionally dispenable genes in each group were counted and bar charts were plotted.

### 3.4.2 Map Conditionally Essential and Dispensable Genes into Pathways

To study functional of essential and dispensable genes with high MFVs (top 10%) and low MFVs (bottom 10%), we mapped these genes into KEGG metabolic pathways [44]. Seven generic metabolic pathways, including biosynthesis of secondary metabolites, carbon metabolism, microbial metaboism in diverse environments, fatty acid metabolism, biosynthesis of amino acids, 2-oxocarboxylic acid metabolism, and degradation of aromatic compounds, are used to do this mapping. These pathways

were selected from metabolic overview pathways of *E. coli K-12 MG1655* in KEGG. Gene to pathway mapping was conducted using data mapping tool supplied in KEGG pathway map.

# Chapter 4

# Results

## 4.1 Flux Distribution of *E. coli* Mutants

### 4.1.1 Flux Distribution Representation

With the minimum and maximum fluxes calculated, we tested best representation of fluxes. Given a growth rate, the flux that can pass through a certain reaction is constrained by a range, between minimum flux and maximum flux. To know which is the best representation of the fluxes passing through a reaction, we plotted the heatmaps of both minimum flux and maximum flux. Minimum fluxes of most reactions are zero, which do not contain much features for all reactions. In the contrary, maximum fluxes show a good flux fingerprint for different conditions, as shown in Figure 4.1. In our analysis, we utilized maximum fluxes as our target to study flux distributions. The later flux distributions are distribution of maximum fluxes by default.

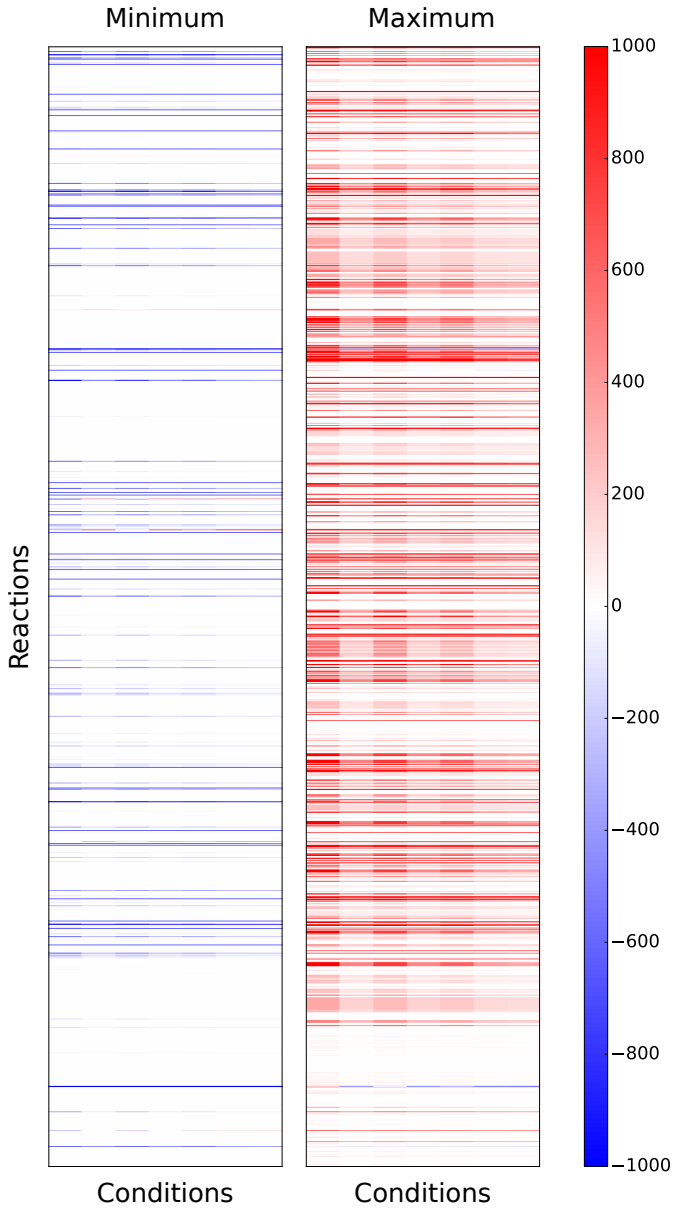


Figure 4.1: Minimum and maximum flux distribution heatmap of the *prpD* mutant. The left heatmap is the minimum fluxes of reactions under 7 termperature conditions, the right is maximum fluxes of reactions under 7 temperature conditions.

### 4.1.2 Flux Distribution under Different Stress Conditions

Figure 4.2 shows distribution of fluxes in three mutants, *thrC*, *tsx*, and *prpD*, under 7 different temperature conditions. Each row represents a reaction and each column represents a tempreature condition. The conditions from left to right are 16℃, 18℃, 20℃, 40℃, 42℃, 43.5℃, and 45℃.

From Figure 4.2, we can see three different flux distribution patterns of mutants over conditions. flux distributions of *prpD* are different in same stress conditions Flux distributions of *tsx*, though change over different conditions, have some similarity in cold conditions (16℃, 18℃, and 20℃) or in heat conditions (40℃, 42℃, 43.5℃, and 45℃). A clear flux distribution similarity of *rfbX* under cold stress conditions are observed, but not in heat stress conditions. These three mutants show different degree of association between flux distribution of a mutant and stress conditions. This discovery is interesting because it shows that there're some patterns in mutant flux distributions under stress conditions.

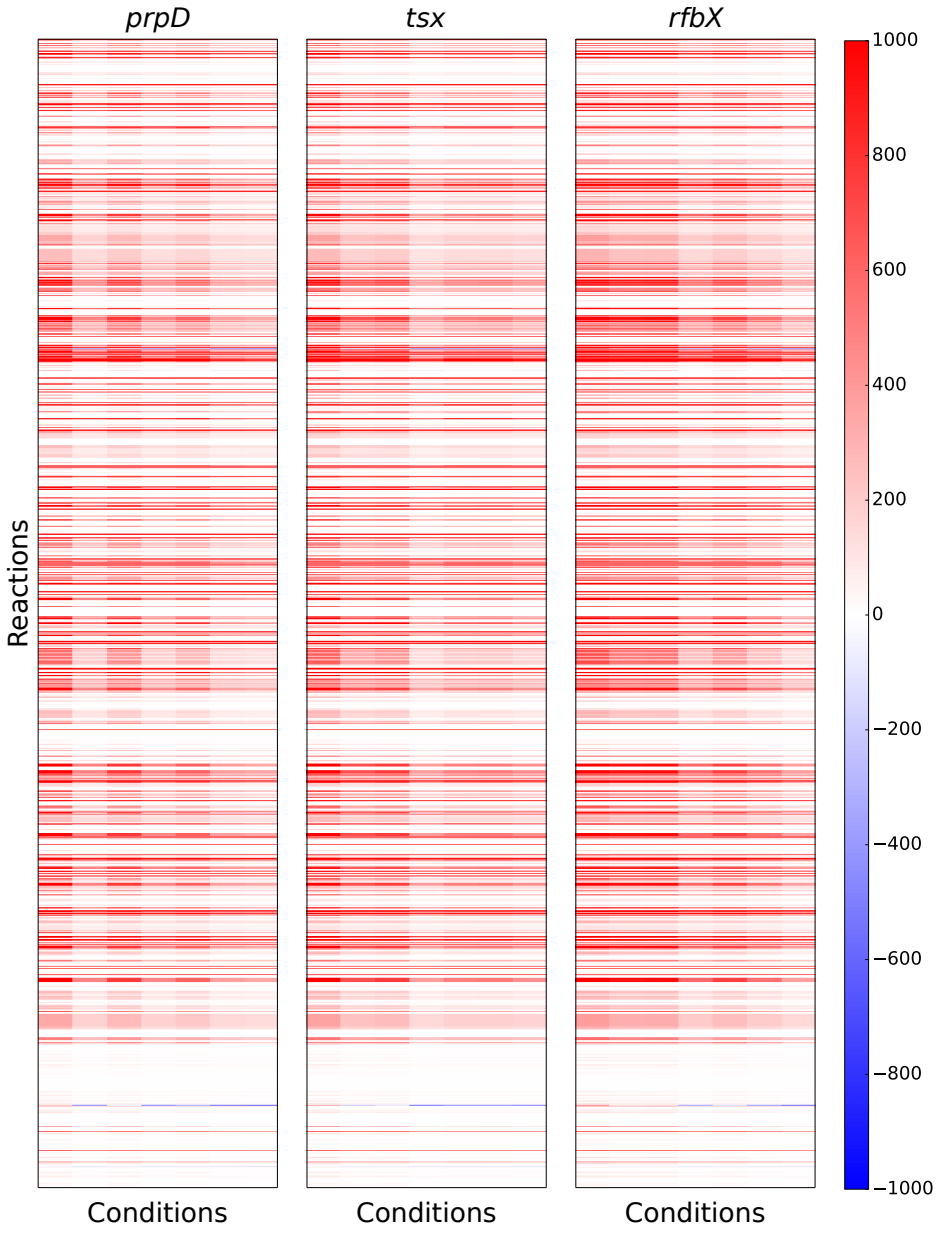


Figure 4.2: Flux distribution heatmap of 3 mutants, *prpD*, *tsx*, and *rfbX*. *prpD* has high variances in fluxes and has no stress condition specificity. *tsx* demonstrates some flux distribution similarity under cold stress conditions and heat stress conditions. *rfbX* shows a clear flux distribution similarity under cold stress conditions.

## 4.2 Flux Variances of the Mutants

To study mutant flux distributions under different stress conditions, we classified these temperature and pH conditions into 4 different stress condition groups based on the underlying *E. coli* stress condition response. The stress condition groups include cold stress conditions, heat stress condition, low pH stress condition, and high pH stress condition. In order to quantatitively analysis mutant flux distributions in stress conditions, we defined a variable MFV to compare between mutants. For each mutant, we calculated variances of fluxes under different conditions in a condition group for all reactions. Average of these variances are taken as MFV of mutant under this condition group.

### 4.2.1 Mutant Flux Variance Distribution under Temperature Stress Conditions

For temperature condition, we calculated the MFV under cold stress conditons and heat stress conditions for mutants. The results are ploted in two histograms as shown in Figure 4.3 A and Figure 4.3 B. In cold stress conditions, majority of MFVs are between 0 and 10000, and few mutants have MFVs scattered between 10000 and 30000. In heat stress conditions, majority of MFVs are between 0 and 4000, and few mutants have MFVs scattered between 4000 and 12000. There are two major peaks in MFV distribution in cold conditions, one is around 0, and the other is around 2500. However, one peak is observed in MFV distribution in heat conditions, which is around 0.

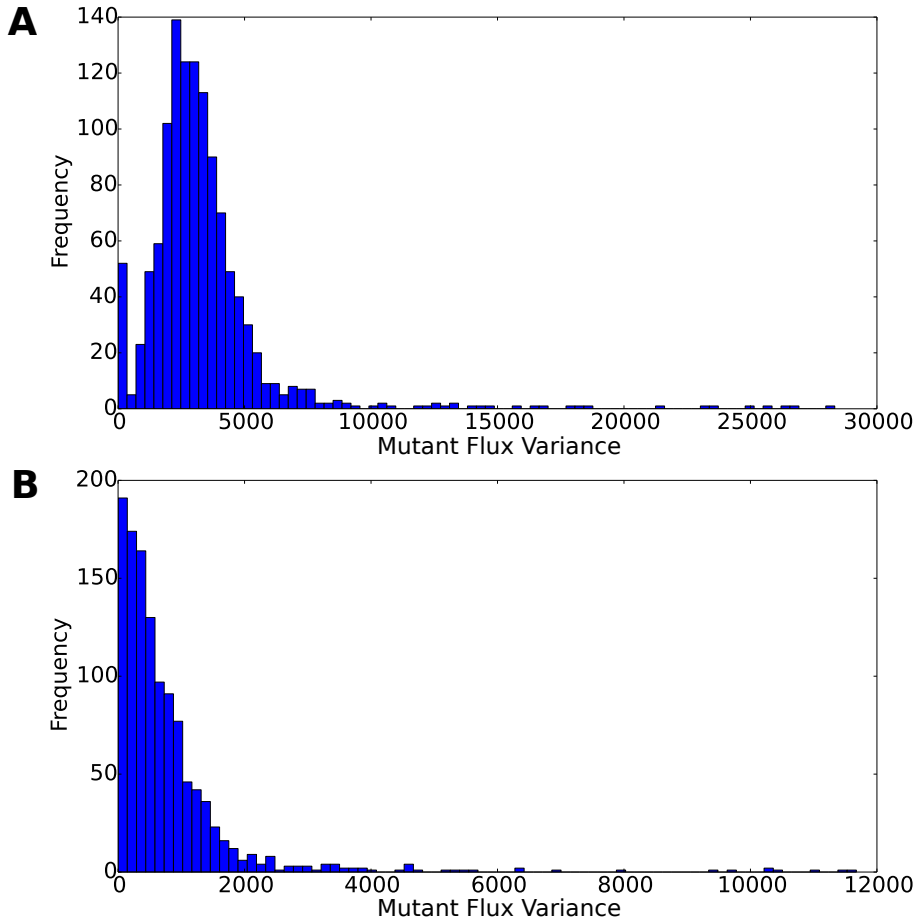


Figure 4.3: MFV histogram for temperature stress condition groups. **A** is the MFV histogram in cold condition group and **B** is MFV histogram in heat condition group.

### 4.2.2 Mutant Flux Variance Distribution under pH Stress Conditions

For pH condition, MFVs under high pH conditons and low pH conditions are calculated for mutants. The results are plotted in two histogram as shown in Figure 4.4 A and Figure 4.4 B. High pH conditons and low pH conditions have similar MFV distribution patterns, there are two peaks in their distribution and majority of them are distributed between 0 and 15000. However, there are some differences between them, the major peak, which centers arround 8000, are fatter in low pH conditions than in high pH conditions, and more mutants have high MFVs in high pH conditions than low pH conditions.

### 4.2.3 Hypotheis based on Mutant Flux Variance Distribution under Stress Conditions

The MFV histograms of 4 condition groups also have some similarity. In all four histograms, a peak around 0 is observed. This is interesting because genes with peak around 0 means that the flux distributions of corresponding mutants do no change over a set of stress condition. That is, the flux distributions of these mutants are robust to stress conditions. Two types of genes might have lead to the flux distribution robustness of mutants to stress conditions. First, some mutated genes might have significant influences on the growth of mutants with the corresponding stress, that metabolism of a gene mutant, growing under conditions in the stress condition group, tend to be similar, Second, some genes might have minor influences on growth under the corresponding stress, that metabolism of a gene mutant under all these stress conditions tends to be similar. The first kind of genes are conditionally essential genes and the second kind of genes are conditionally dispensable genes.

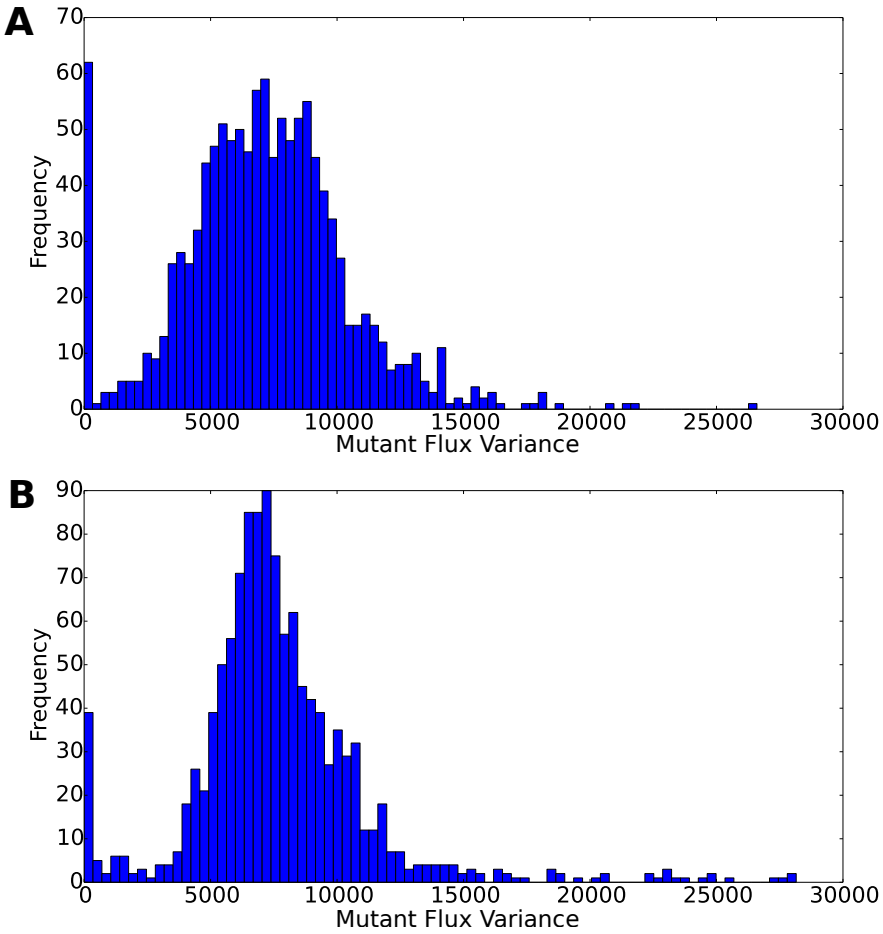


Figure 4.4: MFV histogram for pH stress condition groups. **A** is the MFV histogram in low pH condition group and **B** is MFV histogram in high pH condition group.

## 4.3 Correlation between Gene Conditionally Essentiality and Mutant Flux Variances

### 4.3.1 Definition of Conditionally Essential and Dispensable Genes

In order to test our hypothesis, we defined conditionally essential and conditionally dispensable genes based on experimental data. For each condition group, we defined

conditionally essential genes and conditionally dispensable genes for each condition group based on fitness data. As fitness data are normalized under each condition, the distribution of fitness values under each condition can be assumed to follow a standard normal distribution. Assmuing that fitness values of a mutant are independent with conditions in a stress condition group. Probability of having a mutant with fitness values smaller than -1 or larger than 1 in a stress condition group is very small, (4e-3 for 3 conditions and 6e-4 for 4 conditions, both is smaller than 0.01). If a mutant havs fitness values larger than 1 or smaller than -1 in all conditions within a stress condition group, then the fitness values of this mutant are not conditionally independent with this stress.

Dispensable genes are the genes, which are dispensable in standard laborotary conditions. If mutant of a dispensable gene has high fitness values over all conditions in a stress conditon group, then this gene is defined as a conditionally dispensable gene under this stress. Because with mutation of this gene, *E. coli* can still grow relatively well comparing to other mutants in all conditions of this stress condition group, this gene is relatively dispensable for *E. coli* under this stress. If mutant of a dispensable gene has small fitness values in all conditions in a stress conditon group, then this gene is defined as a conditionally essential gene under this stress. Because with mutation of this gene, growth of *E. coli* is seriously affected in this stress condition group.

### 4.3.2 Gene Conditionally Essentiality and Mutant Flux Variances in Temperature Stress Conditions

We divided mutant corresponding genes into 10 groups, G1 to G10, based on the MFVs. Each group Gi contains genes with MFVs in percentile range between $10(i-1)$ and $10i$. For example, G1 contains genes with MFVs in bottom 10 percent, and G10 contains genes with MFVs in percentile range between 90 and 100. For each bin, we

check how many conditionally essential genes and conditionally dispensable genes are there in it. The fractions of conditionally essential genes and conditionally dispensable genes in each group are plotted as bar charts in Figure 4.5 and Figure 4.6.

| Counts | Cold | | | | Heat | | | |
|---|---|---|---|---|---|---|---|---|
| Group | essential | dispensable | others | Total | essential | dispensable | others | Total |
| G1 | 27 | 11 | 80 | 118 | 7 | 4 | 107 | 118 |
| G2 | 7 | 0 | 110 | 117 | 11 | 1 | 105 | 117 |
| G3 | 0 | 0 | 117 | 117 | 6 | 2 | 110 | 118 |
| G4 | 0 | 0 | 117 | 117 | 1 | 5 | 111 | 117 |
| G5 | 0 | 0 | 118 | 118 | 9 | 1 | 108 | 118 |
| G6 | 0 | 0 | 117 | 117 | 9 | 2 | 106 | 117 |
| G7 | 1 | 0 | 116 | 117 | 5 | 2 | 110 | 117 |
| G8 | 0 | 1 | 116 | 117 | 9 | 3 | 106 | 118 |
| G9 | 0 | 10 | 107 | 117 | 4 | 3 | 110 | 117 |
| G10 | 0 | 45 | 72 | 117 | 0 | 10 | 107 | 117 |
| Total | 35 | 67 | 1070 | 1172 | 61 | 33 | 1080 | 1174 |

Table 4.1: Distribution of essential and dispensable genes in temperature conditions (The essential genes are defined based on fitness values under stress conditons. If a gene has fitness values smaller than -1 in all conditions of a stress condition group, then it is considered a conditionally essential gene. If a gene has fitness values greater than 1 in all conditions of a stress condition group, then it is considered a conditionally dispensable gene. The groups, G1 to G10, are decided by mutant MFVs. Each group Gi contains genes with MFVs in percentile range between $10(i-1)$ and $10i$.)

| P-value | Cold | | Heat | |
|---|---|---|---|---|
| Group | essential | dispensable | essential | dispensable |
| G1 | 1.51e-16 | 0.108 | 0.676 | 0.577 |
| G2 | 0.091 | 0.002 | 0.056 | 0.269 |
| G3 | 0.053 | 0.002 | 1.000 | 0.777 |
| G4 | 0.053 | 0.002 | 0.033 | 0.389 |
| G5 | 0.054 | 0.002 | 0.213 | 0.269 |
| G6 | 0.053 | 0.002 | 0.210 | 0.776 |
| G7 | 0.271 | 0.002 | 0.835 | 0.776 |
| G8 | 0.053 | 0.015 | 0.213 | 1.000 |
| G9 | 0.053 | 0.226 | 0.531 | 1.000 |
| G10 | 0.053 | 1.02e-25 | 0.003 | 0.002 |

Table 4.2: P-values of gene distribution in temperature conditions (P-values are calculated using two-tailed binominal distribution)

For temperature stress conditions, we can clearly see the distribution patterns in cold stress conditions and in heat stress conditions. The frequencies of conditionally

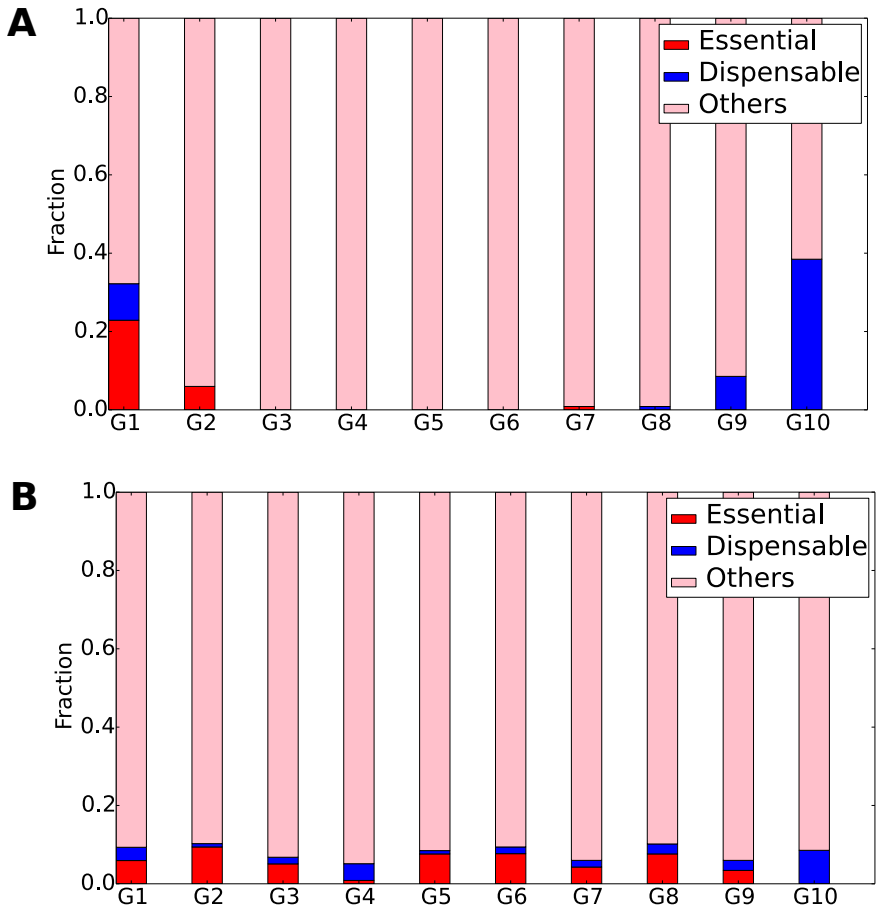


Figure 4.5: Bar charts of conditionally essential genes and conditionally dispensable genes in temperature condition groups. **A** is for cold temperature condition group and **B** is for heat temperature group.

essential genes and conditionally dispensable genes in gene groups are shown in Table 4.1. P-values calculated based on bionominal distribution are shwon in Table 4.2. As shown in Figure 4.5 A, for cold conditions, cold essential genes tend to have small MFVs, and dispensable genes have relatively divergent MFVs, some of them have small MFVs, some of them have large MFVs. Based on the p-values calculated, 27 cold essential genes are in G1 (p-value is 1.51e-16). 45 cold dispensable genes are observed in G10 (p-value is 1.02e-25). Also cold dispensable genes are do not have

MFVs in middle range, from G2 to G7. These discoveries show that cold essential genes tend to have small MFVs, and cold dispensable genes tend to have large MFVs or small MFVs.

In heat conditions, heat essential and dispensable genes tend to be more uniformly distributed, as shown in Figure 4.5 B. However, no heat essential genes are observed in G10 (p-value is 0.003). 10 heat dispensable genes are found in G10 (p-value is 0.002) These observation shows that heat essential genes tend not to have high MFVs and many heat dispensable genes have high MFVs, which is similar to our observation in cold stress conditions.

### 4.3.3 Gene Conditionally Essentiality and Mutant Flux Variances in pH Stress Conditions

| Counts | Low PH | | | | High PH | | | |
|---|---|---|---|---|---|---|---|---|
| Group | essential | dispensable | others | Total | essential | dispensable | others | Total |
| G1 | 16 | 0 | 102 | 118 | 15 | 7 | 97 | 119 |
| G2 | 9 | 0 | 109 | 118 | 11 | 0 | 107 | 118 |
| G3 | 5 | 1 | 112 | 118 | 4 | 0 | 114 | 118 |
| G4 | 6 | 0 | 111 | 117 | 0 | 0 | 118 | 118 |
| G5 | 2 | 1 | 115 | 118 | 0 | 0 | 118 | 118 |
| G6 | 2 | 0 | 116 | 118 | 0 | 0 | 118 | 118 |
| G7 | 1 | 0 | 116 | 117 | 0 | 0 | 118 | 118 |
| G8 | 0 | 0 | 118 | 118 | 0 | 0 | 118 | 118 |
| G9 | 1 | 1 | 116 | 118 | 0 | 6 | 112 | 118 |
| G10 | 0 | 1 | 116 | 117 | 0 | 12 | 106 | 118 |
| Total | 42 | 4 | 1131 | 1177 | 30 | 25 | 1126 | 1181 |

Table 4.3: Distribution of essential and dispensable genes in pH conditions

In pH stress conditions, clear distribution patterns of conditionally essential genes are also observed. As shown in Figure 4.6 A, in low pH conditions, low pH essential genes are more widely distributed. 16 low pH essential genes are observed in G1 (p-value is 5.06e-6). However, only 4 low pH dispensable genes are observed. As the number of metabolic genes is limited, the distribution of low pH dispensable genes is not significant.

In high pH stress conditions, high pH essential genes tend to have small MFVs, and many high pH dispensable genes have large MFVs, as shown in Figure 4.6 B. G1 contains 15 high pH essential genes (p-value is 4.05e-7) and G2 contains 11 high pH essential genes ( p-value is 2.24e-4). Also in G10, 12 high pH dispensable genes are observed ( p-value is 8.65e-6).

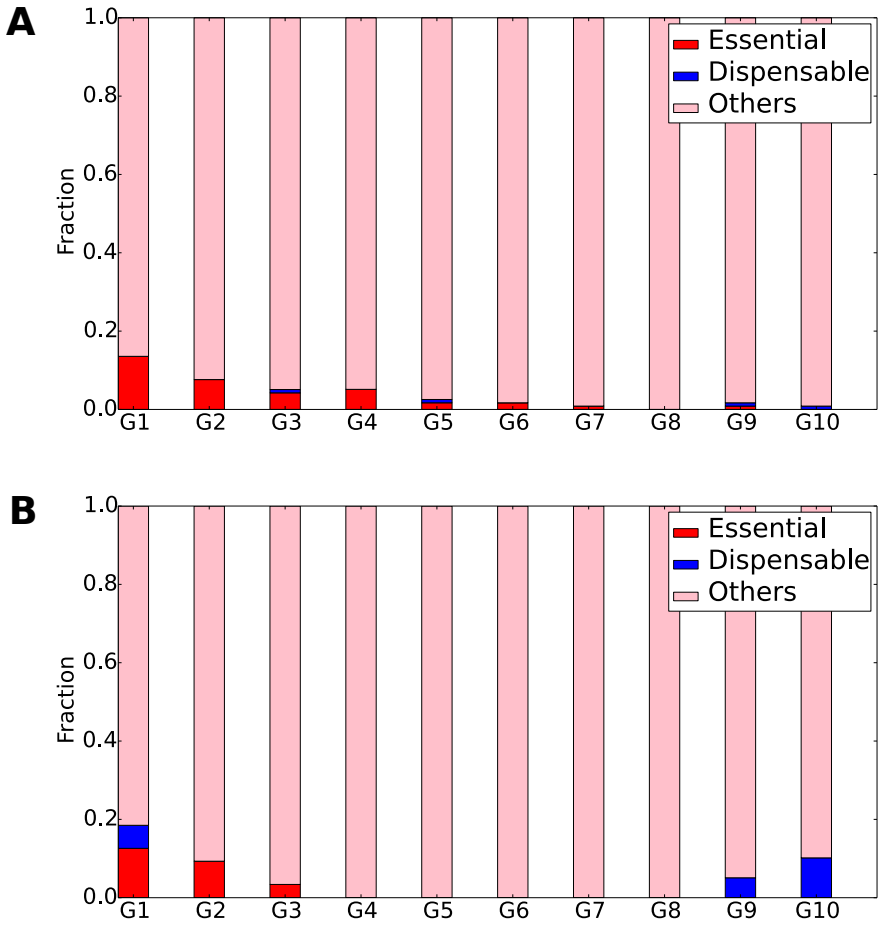


Figure 4.6: Bar charts of conditionally essential genes and conditionally dispensable genes in each MFV group. **A** is for low pH stress conditions. **B** is for high pH stress conditions.

| P-value | Low PH | | High PH | |
|---|---|---|---|---|
| Group | essential | dispensable | essential | dispensable |
| G1 | 5.06e-06 | 1.000 | 4.05e-07 | 0.014 |
| G2 | 0.040 | 1.000 | 2.24e-04 | 0.186 |
| G3 | 0.617 | 0.331 | 0.548 | 0.186 |
| G4 | 0.316 | 1.000 | 0.079 | 0.186 |
| G5 | 0.450 | 0.331 | 0.079 | 0.186 |
| G6 | 0.450 | 1.000 | 0.079 | 0.186 |
| G7 | 0.135 | 1.000 | 0.079 | 0.186 |
| G8 | 0.024 | 1.000 | 0.079 | 0.186 |
| G9 | 0.135 | 0.331 | 0.079 | 0.040 |
| G10 | 0.024 | 0.329 | 0.079 | 8.65e-06 |

Table 4.4: P-values of gene distribution in pH conditions

### 4.3.4 Discovery from Gene Conditionally Essentiality and Mutant Flux Variances Correlation

In summary, distributions of conditionally essential genes and conditionally dispensable genes show that conditionally essential genes tend to have small MFVs, and conditionally dispensable genes tend to have either small or large MFVs. These observations support our hypothesis that conditionally essential genes and conditionally dispensable genes tend to have low MFVs. However, one exception is also observed. Conditionally dispensable genes not only can have small MFVs, but also can have large MFVs. This may indicate that there are two different kinds of metabolic effects associated with conditionally dispensable genes.

## 4.4 Conditionally Essential and Dispensable Genes in Generic Pathways

In previous analysis, we find clear distribution patterns of conditionally essential genes and conditionally dispensable genes in their fluxes. We are interested in these distribution patterns and want to locate these conditionally essential genes and conditionally dispensable genes with either large or small MFVs in generic pathways. *E. coli K-12 MG1655* metabolic overview pathways in KEGG are used to do this pathway enrichment analysis. The pathway distribution of conditionally essential genes and conditionally dispensable genes in gene group G1 and G10 are shown in Table 4.5 and Table 4.6.

### 4.4.1 Temperature Conditionally Essential and Dispensable Genes in Generic Pathways

Examples of temperature essential gene mappings in *E. coli K-12 MG1655* metabolic pathways are shown in Figure 4.7 and Figure 4.8. From these pathway mappings, we see that these genes are associated with different pathways. So we further mapped these conditionally essential genes in G1 and conditionally dispensable genes in G1 and G10 into several generic pathways to see the difference in involvement of pathways, as shown in Table 4.5. Table 4.5 shows the distribution of temperature stress essential and dispensable genes in several generic pathways. From Table 4.5, we can see that these cold and heat essential genes have quite different pathway associations. Cold conditionally essential genes tend to be clustered in biosynthesis of secondary metaboites, while many heat conditionally essential genes are more involved in microbial metabolism in diverse environments. Dispensable genes from G1 and G10 in cold stress, though have highly different MFVs, tend to involve in similar generic pathways.

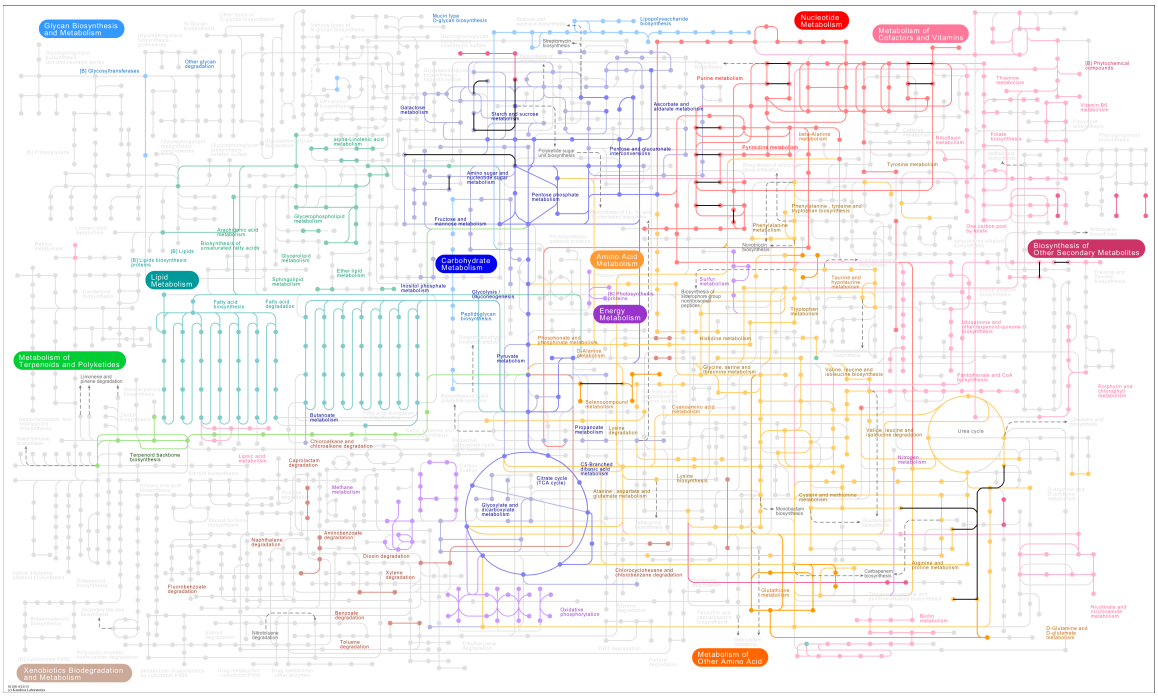


Figure 4.7: Small MFV cold essential genes in *E. coli K-12 MG1655* metabolic pathways. The associated reactions of these genes are marked as black.

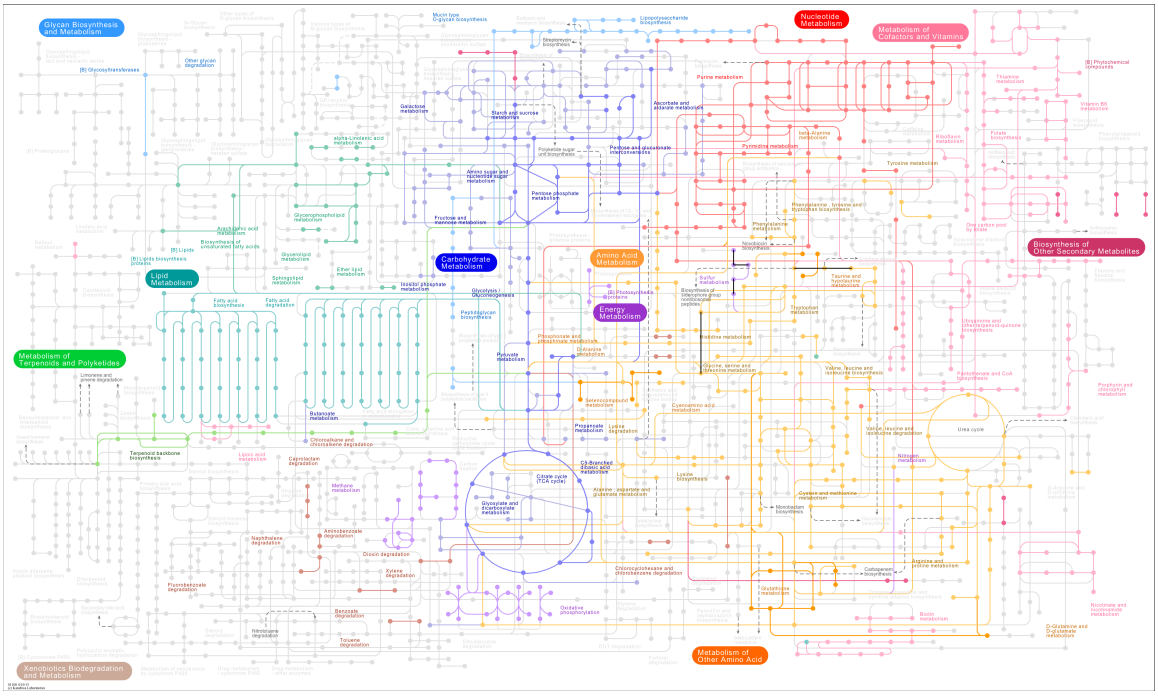


Figure 4.8: Small MFV heat essential genes in *E. coli K-12 MG1655* metabolic pathways. The associated reactions of these genes are marked as black.

| Generic pathways | Essential genes in G1 | | Dispensable genes in G1 | | Dispensable genes in G10 | |
|---|---|---|---|---|---|---|
| | Cold | Heat | Cold | Heat | Cold | Heat |
| Biosynthesis of secondary metabolites | *cpsG,ndk, glgA,glgC, rffH,hemX, hemN* | *trpD* | *gltA,aspC, accD* | *pheA,glmS* | *acnB,ispA, ubiF,sdhB, ispE,gcvH* | *ubiF,sdaA, ubiX,ubiH, fbp* |
| Microbial metabolism in diverse environments | *fucI* | *cysC,cysN, cysD,cysJ, cysE,cysQ* | *gltA,accD, cysD,cysI, cysE* | | *acnB,mhpB, sdhB,aceA* | *mhpB,fbp* |
| Biosynthesis of antibiotics | *ndk,rffH* | *cysQ* | *gltA,aspC, accD* | *pheA,glmS* | *acnB,ispA, sdhB,ispE* | *sdaA,fbp* |
| Carbon metabolism | | *cysE* | *gltA,accD, cysE* | | *acnB,sdhB, gcvP,aceA* | *sdaA,fbp* |
| 2-Oxocarboxylic acid metabolism | | | *gltA,aspC* | | *acnB* | |
| Fatty acid metabolism | | | *accD* | | *fabH* | |
| Biosynthesis of amino acids | | *trpD,cysE* | *gltA,aspC, cysE* | *pheA* | *acnB* | *sdaA* |
| Degradation of aromatic compounds | | | | | *mhpB* | *mhpB* |

Table 4.5: Distribution of temperature conditionally essential genes and conditionally dispensable genes over generic pathways

### 4.4.2 pH Conditionally Essential Genes and Conditionally Dispensable Genes in Generic Pathways

Examples of pH essential gene mappings in *E. coli K-12 MG1655* metabolic pathways are shown in Figure 4.9 and Figure 4.10. Similarly, these genes are associated with different pathways. So we mapped these conditionally essential genes in G1 and conditionally dispensable genes in G1 and G10 into several generic pathways to analyze the difference in involvement of pathways Table 4.6 shows the distribution of pH stress conditionally essential genes and conditionally dispensable genes in several generic pathways. Conditionally essential genes in low pH and high pH involved in similar pathways, except one gene *gltA*, which involved in 2-oxocarboxylic acid metabolism. None of low pH conditionally dispensable genes are mapped into these generic pathways as only one low pH conditonal dispensable gene has MFV in top 10% and botom 10%. This shows that many metabolic genes involved in low pH stress condition response. None of high pH dispensable genes with low MFVs is mapped into these generic pathways.

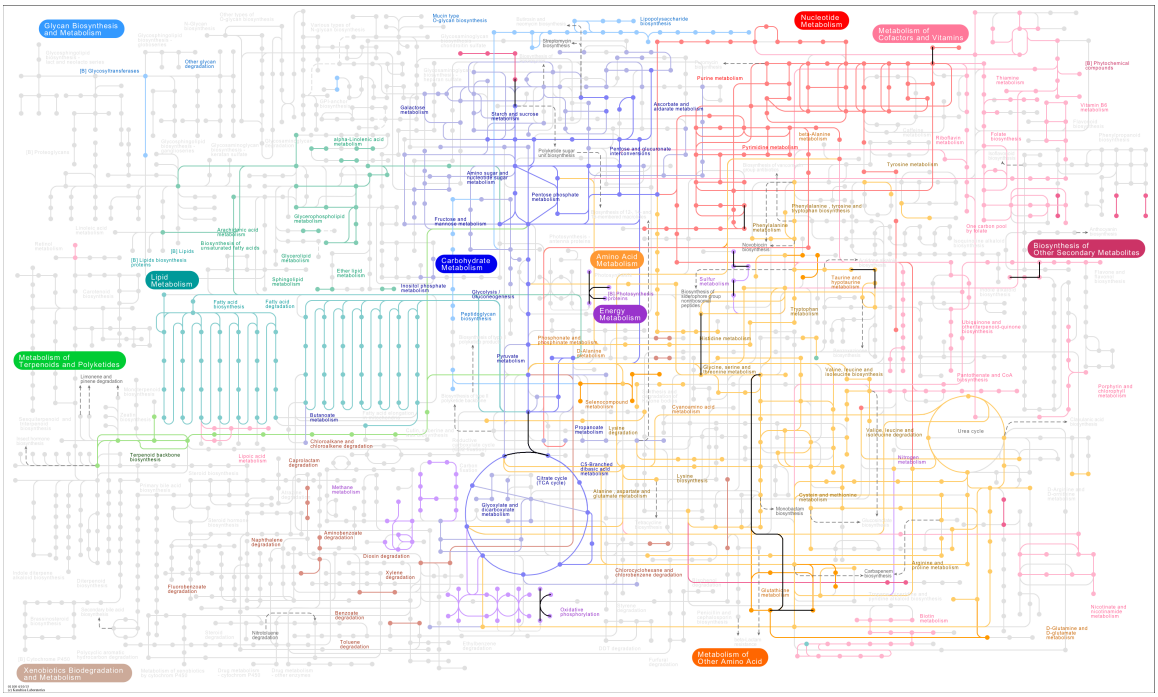


Figure 4.9: Small MFV low pH essential genes in *E. coli K-12 MG1655* metabolic pathways. The associated reactions of these genes are marked as black.

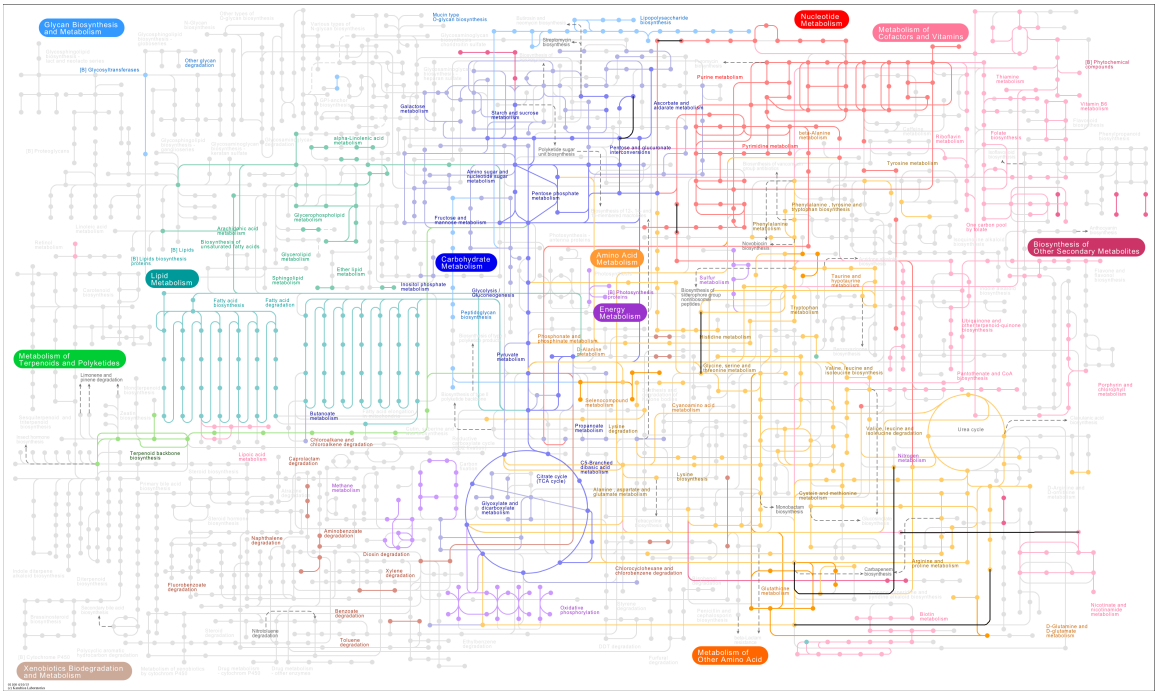


Figure 4.10: Small MFV high pH essential genes in *E. coli K-12 MG1655* metabolic pathways. The associated reactions of these genes are marked as black.

| Generic pathways | Essential genes in G1 | | Dispensable genes in G1 | | Dispensable genes in G10 | |
|---|---|---|---|---|---|---|
| | Low pH | High pH | Low pH | High pH | Low pH | High pH |
| Biosynthesis of secondary metabolites | *gltA,trpC, cysG,rffH* | *entB,purD, idnK* | | | | *glgP,rpiB* |
| Microbial metabolism in diverse environments | *gltA,cysC, cysD,cysI, cysJ,cysE, cysQ* | *cysE,glnA, idnK* | | | | *hyaB,nirB, rpiB* |
| Biosynthesis of antibiotics | *gltA,rffH, cysQ* | *entB,purD, idnK* | | | | *rpiB* |
| Carbon metabolism | *gltA,cysE* | *cysE,idnK* | | | | *rpiB* |
| 2-Oxocarboxylic acid metabolism | *gltA* | | | | | |
| Fatty acid metabolism | | | | | | |
| Biosynthesis of amino acids | *gltA,trpC, cysE* | *cysE,glnA* | | | | *rpiB* |
| Degradation of aromatic compounds | | | | | | |

Table 4.6: Distribution of pH conditionally essential genes and conditionally dispensable genes over generic pathways

## 4.5 Conditionally Essential Genes with Small Mutant Flux Variances

In our study, we retrieved a list of conditionally essential metabolic genes, which have small MFVs, under each sttress conditon group. The essential genes with MFVs in bottom 10% for each condition group are shown in Table 4.7.

| Stress Condition | Genes |
|---|---|
| Cold | *speE, metN, mak, grxA, dadX, sapD, znuA, rfbX, wcaK, cpsG, gmm, ndk, srlA, srlE, fucI, speB, bacA, acrE, frlD, glgA, glgC, malS, rffH, hemX, corA, hemN, yjjL* |
| Heat | *trpD, cysC, cysN, cysD, cysJ, cysE, cysQ* |
| Low pH | *gltA, moaC, trpC, ptsI, cysC, cysD, cysI, cysJ, thyA, gshB, cysG, selA, cysE, atpA, rffH, cysQ* |
| High pH | *nhaA, entB, nadA, poxB, pyrD, nudJ, puuD, gudX, cysE, rfaP, yidK, glnA, purD, treC, idnK* |

Table 4.7: Conditionally essential genes with small MFVs

We searched literature to find direct evidence of these genes involving in *E. coli* stress condition response. For high pH and low pH conditions, there are known metabolic mechanisms related to them. In low pH condition, $F_1F_0$-ATPase plays a very important role [14]. Conditionally essential gene *atpA* discovered in our analysis produces $\alpha$ subunit of $F_1F_0$-ATPase, which fits with this known mechanism. Gene *nhaA*, which produces a $Na^+/H^+$ anti-porter, is known to be important in high pH condition [15]. In our results, gene *nhaA* is discovered to be an essential gene in high pH condition, and have the smallest MFV among all other conditionally essential metabolic genes. These two evidence show that the small MFV conditionally essential metabolic genes discovered by our analysis have a good overlapping with experimental stress conditon essential genes. The conditionally essential metabolic genes, which shown small MFVs, can be experimental targets to study metabolic mechanisms of *E. coli* stress condition response.

We also check the functions of the genes discovered in our study to find potential

correlations with known stress condition mechanisms. The functions and potential relationships are discussed in Section 5.1.

# Chapter 5

# Discussion

## 5.1 Essential Metabolic Genes under Stress Conditions

In our analysis, a set of conditionally essential metabolic genes were found, which have small MFVs. In low pH stress conditions, 16 metabolic genes are discovered to be conditionally essential with small MFVs. *AtpA*, which produces ATP synthase $F_1$ $\alpha$ subunit is include in these 16 metabolic genes. ATPase complex is considered to be important for pH homeostasis in low pH condition. Some other *atp* genes, including *atpF*, *atpB*, and *atpG* are conditionally leathal under some of the low pH condition. The discovery of ATPase complex related genes shows a strong evidence for our methods. In low pH stress, consumption of $H^+$ inside cytoplasm is increased. Low MFV conditionally essential gene *trpC*, *cysI*, *cysD*, and *rffH* are found to catalize metabolic reactions which consumes $H^+$, which fits with $H^+$ consumption mechanism in low pH stress response [45]. Besides, 7 *cys* genes are also found among these conditionally essential genes. *CysC* and *cysD* are known to be regulated by a same protein. *CysI* and *cysJ* are regulated by a same protein. This indicates that these genes have similarities during regulation, thus may have related functions in metabolism of *E. coli* in low pH conditions.

For high pH stress conditions, 14 metabolic genes are discovered to be conditionally essential and have small MFVs. *NhaA*, which produces a $Na^+/H^+$ anti-porter located in the inner-membrane of *E. coli*, is among these genes. It is known that *nhaA* plays a pivotal role in *E. coli*'s adaptation to high pH conditions. *NhaA* imports $H^+$ while exports $Na^+$ out of cytoplasm. Many other genes, including *poxB*, *pyrD*, *rfaP*, *nudJ*, *glnA*, *purD*, and *idnK*, are found to catalize reactions which produces $H^+$. $H^+$ production inside cytoplasm is an important function for *E. coli* to maintain a pH homestasis when exposed in high pH stress conditions. These metabolic genes may involved in increased $H^+$ production of *E. coli* under high pH conditions. Also, *yidK*, which produces a $Na^+$ driven metabolite uptake transporter, is found. Working together with *nhaA*, *yidK* may balances the cytoplasmic $Na^+$ while *nhaA* imports more $H^+$ into the cytoplasm during high pH stress conditions.

Current known mechanisms for temperature stress response are about transcription regulation. So we tried to overlap known repressed/induced genes in heat and cold shock response with the metabolic genes we discovered. In cold stress condition, 27 metabolic genes are discovered to be conditionally essential with small MFVs. Among these genes, *srlA* is found to be transiently induced and *rffH* is discovered to be transiently repressed during cold shock response of *E. coli* [86]. In heat stress conditions, 7 metabolic genes are discovered to be conditionally essential and have small MFVs, which include *trpD*, *cysC*, *cysN*, *cysD*, *cysJ*, *cysE*, and *cysQ*. None of these genes is induced by $\sigma 32$ [87]. This indicates that metabolic genes required for *E. coli* to survive in heat stress conditions might not directly related to $\sigma 32$, which fits with a previous study in *Yeast* [88].

Stress condition response is a complex process in an organism, which is a combined effect of transcriptional regulation, transcriptional regulation, signaling and metabolism. For temperature stress condition, transcription and translation level mechanisms in *E. coli* are well-studied, but metabolism level changes are not clear

[17, 58, 59, 61, 62]. The metabolic genes from our study might be future experimental candidates to learn metabolic mechanisms for temperature stress condition response. For pH stress condition, some general metabolism level mechanisms are discovered. In our analysis, *atpA* is found to be essential under low pH stress conditions and *nhaA* is found to be essential under high pH stress conditions. Metabolic mechanisms of *E. coli* pH stress response indicates that these two genes are important during low pH stress and high pH stress separately, which gives strong evidences of our computational approach. These evidences drive us to propose a hypothesis that low MFV conditionally essential genes, predicted by our simulation, might be conditionally essential genes. Thus the genes predicted by our method can be future candidates in experimental analysis of *E. coli* under temperature and pH stress conditions.

## 5.2 Metabolic Networks in Stress Condition Analysis

In this study, we combine metabolic networks with traditional gene mutant fitness data to study conditionally essential genes in stress conditions. In general, our method combines the information from fitness of gene knockouts with the metabolic functions of gene products to determine the conditional essentiality of metabolic genes. Traditional experimental study determines the conditional essentiality of a gene by checking the conditional fitness of gene mutants. If knockout of a gene is lethal under some stress conditions, then it is considered to be conditional essential. Some genes, knockout of which cause small growth rates, are ignored by these studies [8]. Our method can find some essential genes which show small growth rates.

Previous computation investigations of stress conditionally essential genes are more focused on carbon substrate stress conditions [2, 78, 89]. These studies give us some essential genes involved in carbon substrate stress conditions. Our study

combines both metabolic functions of genes and mutant stress condition fitness of genes, and can expand our knowledge of conditionally essential genes in non-carbon stress conditions.

# Chapter 6

# Summary and Future Work

## 6.1 Summary

We used FBA to analyze flux distributions of mutants under stress conditions. CBMs of 1258 mutants were constructed based on *E. coli* metabolic model iJO1366. Protein abundance data were used during construction of mutant CBMs. FBA was conducted for each mutant CBM to get the optimal growth rate of the mutant. Then, mutant fitness data under different stress conditions were mapped into growth rate data based on mutant optimal growth rates. After that, FVA was conducted for each of mutant CBM under each stress condition with flux of biomass reaction set as the corresponding growth rate. From FVA, fluxes distribution of each mutant under each stress condition was retrieved.

We analyzed the distribution of fluxes under specific condition groups based on the stress condition effects. Four condition groups were analyzed in our study, which include cold stress conditions, heat stress conditions, low pH stress conditions, and high pH stress conditions. A value, MFV, was calculated for each mutant in each stress condition group. Mutant MFVs were used to compare the mutants under each stress condition group. Distributions of mutant MFV in each stress condition group were plotted as a histogram. These histograms have a peak around 0.

To analyze corresponding genes distribution in this peak, we defined condition-

ally essential and dispensable genes for each conditional group. These conditionally essential and dispensable genes are dispensable genes for *E. coli*, which are defined in standard laborotary conditions. Dispensable genes, whose corresponding mutants have fitness values smaller than -1 in all conditions of a stress condition group, are defined as conditionally essential genes in this stress condition. Dispensable genes, whose corresponding mutants have fitness values larger than 1 in all conditions of a stress condition group, are defined as conditionally dispensable genes in this stress condition. We checked the location of conditionally essential genes and conditionally dispensable genes in the MFV distribution histogram by plotting a bar chart for each stress condition group. The bar chart was plotted by dividing the MFVs into 10 groups, with each group containing 10% of mutants.

We found that conditionally essential genes tend to have small MFVs, and conditionally dispensalbe genes tends to have either small or large MFVs. We checked the conditionally essential genes with MFVs in bottom 10%, and found experimental evidence of conditional essentiality. *AtpA* is found to subunit of ATPase, which has important role under low pH stress conditions. *AhaA* is known to be essential under high pH stress conditions. We also checked functions and relationships of other genes with their corresponding stress conditions. The metabolic genes discovered in our analysis might be potential targets for experimental study.

## 6.2 Future Research Work

In this study, we found some *E. coli* conditionally essential genes in each stress condition group. We checked existing literatues about *E. coli* stress condition analysis to find essentiality evidence for these genes. In future analysis, we want to conduct sequence similarity search using sequences of discovered genes in our study against all genes in some other organisms, such as *Yeast*, which may enable us find conditionally

essential genes in other species or find conditional essentiality evidence of the genes discovered by us.

In this study, we checked conditionally essential genes with low MFVs. In future analysis, we also want to study conditionally dispensable genes with MFVs in top 10% and bottom 10%, which is also an interesting discovery in our analysis.

The major focus of this study is about gene-stress condition relationship. In future analysis, we also want to study the relationships between metabolic reactions and stress conditions. This may enable us find some interesting metabolic pathways associated with stress conditions, and some interesting metabolic network topologies which can explain conditional essentiality and dispensability of genes.

# REFERENCES

[1] R. Harrison, B. Papp, C. Pál, S. G. Oliver, and D. Delneri, "Plasticity of genetic interactions in metabolic networks of yeast," *Proceedings of the National Academy of Sciences*, vol. 104, no. 7, pp. 2307–2312, 2007.

[2] S. K. Remold and R. E. Lenski, "Pervasive joint influence of epistasis and plasticity on mutational effects in escherichia coli," *Nature genetics*, vol. 36, no. 4, pp. 423–426, 2004.

[3] Z. Gu, L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W.-H. Li, "Role of duplicate genes in genetic robustness against null mutations," *Nature*, vol. 421, no. 6918, pp. 63–66, 2003.

[4] S. Gerdes, M. Scholle, J. Campbell, G. Balazsi, E. Ravasz, M. Daugherty, A. Somera, N. Kyrpides, I. Anderson, M. Gelfand *et al.*, "Experimental determination and system level analysis of essential genes in escherichia coli mg1655," *Journal of bacteriology*, vol. 185, no. 19, pp. 5673–5684, 2003.

[5] C. D. Herring and F. R. Blattner, "Conditional lethal amber mutations in essential escherichia coli genes," *Journal of bacteriology*, vol. 186, no. 9, pp. 2673–2681, 2004.

[6] Y. Kang, T. Durfee, J. D. Glasner, Y. Qiu, D. Frisch, K. M. Winterberg, and F. R. Blattner, "Systematic mutagenesis of the escherichia coli genome," *Journal of bacteriology*, vol. 186, no. 15, pp. 4921–4930, 2004.

[7] F. Arigoni, F. Talabot, M. Peitsch, M. D. Edgerton, E. Meldrum, E. Allet, R. Fish, T. Jamotte, M.-L. Curchod, and H. Loferer, "A genome-based approach for the identification of essential bacterial genes," *Nature biotechnology*, vol. 16, no. 9, pp. 851–856, 1998.

[8] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, A. Wong *et al.*, "Phenotypic landscape of a bacterial

cell," *Cell*, vol. 144, no. 1, pp. 143–156, 2011.

[9] P. G. Jones and M. Inouye, "The cold-shock response: a hot topic," *Molecular microbiology*, vol. 11, no. 5, pp. 811–818, 1994.

[10] R. A. VanBogelen and F. C. Neidhardt, "Ribosomes as sensors of heat and cold shock in escherichia coli." *Proceedings of the National Academy of Sciences*, vol. 87, no. 15, pp. 5589–5593, 1990.

[11] P. G. Jones, R. A. VanBogelen, and F. C. Neidhardt, "Induction of proteins in response to low temperature in escherichia coli." *Journal of Bacteriology*, vol. 169, no. 5, pp. 2092–2095, 1987.

[12] J. W. Erickson and C. A. Gross, "Identification of the sigma e subunit of escherichia coli rna polymerase: a second alternate sigma factor involved in high-temperature gene expression." *Genes & development*, vol. 3, no. 9, pp. 1462–1471, 1989.

[13] J. L. Slonczewski, B. P. Rosen, J. R. Alger, and R. M. Macnab, "ph homeostasis in escherichia coli: measurement by 31p nuclear magnetic resonance of methylphosphonate and phosphate," *Proceedings of the National Academy of Sciences*, vol. 78, no. 10, pp. 6271–6275, 1981.

[14] P. Lund, A. Tramonti, and D. De Biase, "Coping with low ph: molecular strategies in neutralophilic bacteria," *FEMS microbiology reviews*, vol. 38, no. 6, pp. 1091–1125, 2014.

[15] E. Padan, E. Bibi, M. Ito, and T. A. Krulwich, "Alkaline ph homeostasis in bacteria: new insights," *Biochimica et biophysica acta (BBA)-biomembranes*, vol. 1717, no. 2, pp. 67–88, 2005.

[16] T. A. Krulwich, G. Sachs, and E. Padan, "Molecular aspects of bacterial ph sensing and homeostasis," *Nature Reviews Microbiology*, vol. 9, no. 5, pp. 330–343, 2011.

[17] H. Kurata, H. El-Samad, R. Iwasaki, H. Ohtake, J. C. Doyle, I. Grigorova, C. A. Gross, and M. Khammash, "Module-based analysis of robustness tradeoffs in the heat shock response system," *PLoS computational biology*, vol. 2, no. 7, p. e59, 2006.

[18] H. El-Samad and M. Khammash, "Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks," *Biophysical journal*, vol. 90, no. 10, pp. 3749–3761, 2006.

[19] H. El-Samad, H. Kurata, J. Doyle, C. Gross, and M. Khammash, "Surviving heat shock: control strategies for robustness and performance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 8, pp. 2736–2741, 2005.

[20] R. Srivastava, M. Peterson, and W. Bentley, "Stochastic kinetic analysis of the escherichia coli stress circuit using $\sigma$32-targeted antisense," *Biotechnology and Bioengineering*, vol. 75, no. 1, pp. 120–129, 2001.

[21] J. Weber, F. Hoffmann, and U. Rinas, "Metabolic adaptation of escherichia coli during temperature-induced recombinant protein production: 2. redirection of metabolic fluxes," *Biotechnology and bioengineering*, vol. 80, no. 3, pp. 320–330, 2002.

[22] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection," *Molecular systems biology*, vol. 2, no. 1, 2006.

[23] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan *et al.*, "Interaction network containing conserved and essential protein complexes in escherichia coli," *Nature*, vol. 433, no. 7025, pp. 531–537, 2005.

[24] E. C. Hobbs, J. L. Astarita, and G. Storz, "Small rnas and small proteins involved in resistance to cell envelope stress and acid shock in escherichia coli: analysis of a bar-coded mutant collection," *Journal of bacteriology*, vol. 192, no. 1, pp. 59–67, 2010.

[25] M. Wang, M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner, and C. von Mering, "Paxdb, a database of protein abundance averages across all three domains of life," *Molecular & cellular proteomics*, vol. 11, no. 8, pp. 492–500, 2012.

[26] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and

B. Ø. Palsson, "A comprehensive genome-scale reconstruction of escherichia coli metabolism2011," *Molecular systems biology*, vol. 7, no. 1, 2011.

[27] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[28] F. Horn and R. Jackson, "General mass action kinetics," *Archive for rational mechanics and analysis*, vol. 47, no. 2, pp. 81–116, 1972.

[29] J. E. Dowd and D. S. Riggs, "A comparison of estimates of michaelis-menten kinetic constants from various linear transformations," *J. biol. Chem*, vol. 240, no. 2, pp. 863–869, 1965.

[30] J. N. Weiss, "The hill equation revisited: uses and misuses." *The FASEB Journal*, vol. 11, no. 11, pp. 835–841, 1997.

[31] G. PETTERSSON, "The transient-state kinetics of two-substrate enzyme systems operating by an ordered ternary-complex mechanism," *European Journal of Biochemistry*, vol. 69, no. 1, pp. 273–278, 1976.

[32] W. W. Cleland, "Derivation of rate equations for multisite ping-pong mechanisms with ping-pong reactions at one or more sites," *Journal of Biological Chemistry*, vol. 248, no. 24, pp. 8353–8355, 1973.

[33] N. D. Price, I. Famili, D. A. Beard, and B. Ø. Palsson, "Extreme pathways and kirchhoff's second law." *Biophysical journal*, vol. 83, no. 5, p. 2879, 2002.

[34] S. Schuster and C. Hilgetag, "On elementary flux modes in biochemical reaction systems at steady state," *Journal of Biological Systems*, vol. 2, no. 02, pp. 165–182, 1994.

[35] J. L. Reed, "Descriptive and predictive applications of constraint-based metabolic models," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.* IEEE, 2009, pp. 5460–5463.

[36] I. Thiele and B. Ø. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature protocols*, vol. 5, no. 1, pp. 93–121, 2010.

[37] K. Raman and N. Chandra, "Flux balance analysis of biological systems: applications and challenges," *Briefings in bioinformatics*, vol. 10, no. 4, pp. 435–449,

2009.

[38] K. J. Kauffman, P. Prakash, and J. S. Edwards, "Advances in flux balance analysis," *Current opinion in biotechnology*, vol. 14, no. 5, pp. 491–496, 2003.

[39] S. Ranganathan, P. F. Suthers, and C. D. Maranas, "Optforce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions," *PLoS computational biology*, vol. 6, no. 4, p. e1000744, 2010.

[40] N. E. Lewis and A. M. Abdel-Haleem, "The evolution of genome-scale models of cancer metabolism," *Frontiers in physiology*, vol. 4, 2013.

[41] T. Shlomi, T. Benyamini, E. Gottlieb, R. Sharan, and E. Ruppin, "Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect," *PLoS computational biology*, vol. 7, no. 3, p. e1002018, 2011.

[42] A. Raghunathan, S. Shin, and S. Daefler, "Systems approach to investigating host-pathogen interactions in infections with the biothreat agent francisella. constraints-based model of francisella tularensis," *BMC systems biology*, vol. 4, no. 1, p. 118, 2010.

[43] D. R. Hyduke, N. E. Lewis, and B. Ø. Palsson, "Analysis of omics data with genome-scale models of metabolism," *Mol. BioSyst.*, vol. 9, no. 2, pp. 167–174, 2013.

[44] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in kegg," *Nucleic acids research*, vol. 42, no. D1, pp. D199–D205, 2014.

[45] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp, "Ecocyc: a comprehensive database resource for escherichia coli," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D334–D337, 2005.

[46] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard, "Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox," *Nature protocols*, vol. 2, no. 3, pp. 727–738, 2007.

[47] B. Palsson and K. Zengler, "The challenges of integrating multi-omic data sets,"

*Nature chemical biology*, vol. 6, no. 11, pp. 787–789, 2010.

[48] N. Christian, P. May, S. Kempa, T. Handorf, and O. Ebenhöh, "An integrative approach towards completing genome-scale metabolic networks," *Molecular bioSystems*, vol. 5, no. 12, pp. 1889–1903, 2009.

[49] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. De Longueville, E. S. Kawasaki, K. Y. Lee *et al.*, "The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[50] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[51] J. L. Reed, I. Famili, I. Thiele, and B. O. Palsson, "Towards multidimensional genome annotation," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 130–141, 2006.

[52] A. M. Feist and B. Ø. Palsson, "The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli," *Nature biotechnology*, vol. 26, no. 6, pp. 659–667, 2008.

[53] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature*, vol. 429, no. 6987, pp. 92–96, 2004.

[54] D. R. Hyduke and B. Ø. Palsson, "Towards genome-scale signalling-network reconstructions," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 297–307, 2010.

[55] J. Edwards and B. Palsson, "The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences*, vol. 97, no. 10, pp. 5528–5533, 2000.

[56] J. L. Reed, T. D. Vo, C. H. Schilling, B. O. Palsson *et al.*, "An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr)," *Genome Biol*, vol. 4, no. 9, p. R54, 2003.

[57] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson, "A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information," *Molecular systems biology*, vol. 3, no. 1, 2007.

[58] S. Phadtare and K. Severinov, "Rna remodeling and gene regulation by cold shock proteins," *RNA biology*, vol. 7, no. 6, pp. 788–795, 2010.

[59] C. Bárria, M. Malecki, and C. M. Arraiano, "Bacterial adaptation to cold," *Microbiology*, vol. 159, no. Pt 12, pp. 2437–2443, 2013.

[60] T. Yura, "Regulation and conservation of the heat-shock transcription factor $\sigma$32," *Genes to cells*, vol. 1, no. 3, pp. 277–284, 1996.

[61] E. Guisbert, T. Yura, V. A. Rhodius, and C. A. Gross, "Convergence of molecular, modeling, and systems approaches for an understanding of the escherichia coli heat shock response," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 3, pp. 545–554, 2008.

[62] A. S. Meyer and T. A. Baker, "Proteolysis in the escherichia coli heat shock response: a player at many levels," *Current opinion in microbiology*, vol. 14, no. 2, pp. 194–199, 2011.

[63] J. Lin, I. S. Lee, J. Frey, J. L. Slonczewski, and J. W. Foster, "Comparative analysis of extreme acid survival in salmonella typhimurium, shigella flexneri, and escherichia coli." *Journal of bacteriology*, vol. 177, no. 14, pp. 4097–4104, 1995.

[64] S. Bearson, B. Bearson, and J. W. Foster, "Acid stress responses in enterobacteria," *FEMS microbiology letters*, vol. 147, no. 2, pp. 173–180, 1997.

[65] J. Lin, M. P. Smith, K. C. Chapin, H. S. Baik, G. N. Bennett, and J. W. Foster, "Mechanisms of acid resistance in enterohemorrhagic escherichia coli." *Applied and Environmental Microbiology*, vol. 62, no. 9, pp. 3094–3100, 1996.

[66] D. W. Whitman and A. A. Agrawal, "What is phenotypic plasticity and why is it important," *Phenotypic plasticity of insects*, vol. 10, pp. 1–63, 2009.

[67] T. D. Price, A. Qvarnström, and D. E. Irwin, "The role of phenotypic plasticity in driving genetic evolution," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1523, pp. 1433–1440, 2003.

[68] A. A. Agrawal, "Phenotypic plasticity in the interactions and evolution of species," *Science*, vol. 294, no. 5541, pp. 321–326, 2001.

[69] U. N. Lele, U. I. Baig, and M. G. Watve, "Phenotypic plasticity and effects of selection on cell division symmetry in escherichia coli," *PloS one*, vol. 6, no. 1, p. e14516, 2011.

[70] A. M. Dean, "A molecular investigation of genotype by environment interactions." *Genetics*, vol. 139, no. 1, pp. 19–33, 1995.

[71] D. E. Dykhuizen and D. L. Hartl, "Functional effects of pgi allozymes in escherichia coli," *Genetics*, vol. 105, no. 1, pp. 1–18, 1983.

[72] A. Makhorin, "Gnu linear programming kit, version 4.9. gnu software foundation," 2006.

[73] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "Cobrapy: constraints-based reconstruction and analysis for python," *BMC systems biology*, vol. 7, no. 1, p. 74, 2013.

[74] J. D. Orth and B. Palsson, "Gap-filling analysis of the ijo1366 escherichia coli metabolic network reconstruction for discovery of metabolic functions," *BMC systems biology*, vol. 6, no. 1, p. 30, 2012.

[75] I. Tawornsamretkit, R. Thanasomboon, J. Thaiprasit, D. Waraho, S. Cheevadhanarak, and A. Meechai, "Analysis of metabolic network of synthetic escherichia coli producing linalool using constraint-based modeling," *Procedia Computer Science*, vol. 11, pp. 24–35, 2012.

[76] M. Riley, T. Abe, M. B. Arnaud, M. K. Berlyn, F. R. Blattner, R. R. Chaudhuri, J. D. Glasner, T. Horiuchi, I. M. Keseler, T. Kosuge *et al.*, "Escherichia coli k-12: a cooperatively developed annotation snapshot2005," *Nucleic acids research*, vol. 34, no. 1, pp. 1–9, 2006.

[77] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley,

J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew *et al.*, "The complete genome sequence of escherichia coli k-12," *science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[78] E. S. Snitkin, A. M. Dudley, D. M. Janse, K. Wong, G. M. Church, and D. Segrè, "Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions," *Genome Biol*, vol. 9, no. 9, p. R140, 2008.

[79] N. D. Price, J. L. Reed, and B. Ø. Palsson, "Genome-scale models of microbial cells: evaluating the consequences of constraints," *Nature Reviews Microbiology*, vol. 2, no. 11, pp. 886–897, 2004.

[80] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, "Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells," *Science*, vol. 329, no. 5991, pp. 533–538, 2010.

[81] X. Fu, X. Shi, L. Yan, H. Zhang, and Z. Chang, "Supplementary data file in vivo substrate diversity and preference of small heat shock protein ibpb as revealed by a genetically incorporated photo-crosslinker."

[82] J. C. Wright, M. O. Collins, L. Yu, L. Käll, M. Brosch, and J. S. Choudhary, "Enhanced peptide identification by electron transfer dissociation using an improved mascot percolator," *Molecular & Cellular Proteomics*, vol. 11, no. 8, pp. 478–491, 2012.

[83] F. Mancuso, J. Bunkenborg, M. Wierer, and H. Molina, "Data extraction from proteomics raw data: An evaluation of nine tandem ms tools using a large orbitrap data set," *Journal of proteomics*, vol. 75, no. 17, pp. 5293–5303, 2012.

[84] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation," *Nature biotechnology*, vol. 25, no. 1, pp. 117–124, 2007.

[85] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer *et al.*, "Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models," *Molecular systems biology*, vol. 6, no. 1, 2010.

[86] S. Phadtare and M. Inouye, "Genome-wide transcriptional analysis of the cold shock response in wild-type and cold-sensitive, quadruple-csp-deletion strains of escherichia coli," *Journal of Bacteriology*, vol. 186, no. 20, pp. 7007–7014, 2004.

[87] G. Nonaka, M. Blankschien, C. Herman, C. A. Gross, and V. A. Rhodius, "Regulon and promoter analysis of the e. coli heat-shock factor, $\sigma 32$, reveals a multifaceted cellular response to heat stress," *Genes & development*, vol. 20, no. 13, pp. 1776–1789, 2006.

[88] P. A. Gibney, C. Lu, A. A. Caudy, D. C. Hess, and D. Botstein, "Yeast metabolic and signaling genes are required for heat-shock survival and have little overlap with the heat-induced genes," *Proceedings of the National Academy of Sciences*, vol. 110, no. 46, pp. E4393–E4402, 2013.

[89] J. M. Monk, P. Charusanti, R. K. Aziz, J. A. Lerman, N. Premyodhin, J. D. Orth, A. M. Feist, and B. Ø. Palsson, "Genome-scale metabolic reconstructions of multiple escherichia coli strains highlight strain-specific adaptations to nutritional environments," *Proceedings of the National Academy of Sciences*, vol. 110, no. 50, pp. 20 338–20 343, 2013.