DEVELOPMENT OF SYSTEMATIC AND COMBINATORIAL APPROACHES FOR THE METABOLIC ENGINEERING OF MICROORGANISMS

by

Hal Alper

B.S., Chemical Engineering

University of Maryland, College Park, 2002

Submitted to the Department of Chemical Engineering
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Chemical Engineering

at the

Massachusetts Institute of Technology

June 2006

Signature of Author .....
Hal Alper
Department of Chemical Engineering
April 3, 2006

Certified by .....
Gregory Stephanopoulos
Professor of Chemical Engineering
Thesis Supervisor

Accepted by .....
William Deen
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

DEVELOPMENT OF SYSTEMATIC AND COMBINATORIAL APPROACHES FOR THE METABOLIC ENGINEERING OF MICROORGANISMS

by

Hal Alper

Submitted to the Department of Chemical Engineering on April 3, 2006
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Explorations and optimizations through the genomic space are a daunting undertaking given the complexity and size of the possible search space. To approach this problem, systematic and combinatorial approaches were employed for the engineering of cellular phenotype in *Escherichia coli*. Initially, a computational method based on global cellular stoichiometry was employed to identify single and multiple gene knockout targets for lycopene production in *E. coli*. These targets led to substantial increases in lycopene production, but were limited in scope due to the nature of these models. Therefore, these approaches and targets were complemented with combinatorial searches to identify unknown and regulatory targets. When combined, these searches led to further increases of lycopene production and allowed for the visualization of the resulting metabolic landscape. A more exhaustive search was conducted in the background of eight genotypes which resulted in the formulation of the gene knockout search network. This network enables the investigation into how phenotype optimization is biased by search strategy. Collectively, these results demonstrated that despite the complexity and nonlinearity of genotype-phenotype spaces, most of the significant phenotypes were controlled and regulated by a small subset of key "gateway" nodes. Often, the mutations and genotypes incurred in altering global cellular phenotypes are not necessarily additive and can be quite non-linear.

Effective probing of a metabolic landscape requires not only gene deletions, but also the varying (or tuning) of expression level for a gene of interest. Through promoter engineering, a library of promoters of varying strength were obtained through mutagenesis of a constitutive promoter. A multi-faceted characterization of the library, especially at the single-cell level to ensure homogeneity, permitted quantitative assessment correlating the effect of gene expression levels to improved growth and product formation phenotypes in *E. coli*. Integration of these promoters into the chromosome can allow for a quantitative, accurate assessment and tuning of genetic control. Collectively, quantitative phenotype-genotype analysis illustrated that optimal gene expression levels are variable and dependent on the genetic background of the strain.

As a result, tools such as promoter engineering, which allow for a wide range of expression levels, constitutes an integral platform for functional genomics, synthetic biology, and metabolic engineering endeavors.

Finally, multiple genetic modifications are necessary to unlock latent cellular potential. However, the capacity to make these meaningful modifications has remained an elusive task for cellular and metabolic engineering. The tool of global Transcription Machinery Engineering (gTME) allows one to explore a vastly unexplored, expanded search space in a high throughput manner by evaluating multiple, simultaneous gene alterations in order to improve complex cellular phenotypes. Through the alteration of key proteins involved in global transcription, cells may be reprogrammed for phenotypes of interest. Results in phenotype optimization using gTME outperformed traditional approaches to these problems, exceeding, *in a matter of weeks*, benchmarks achieved through decades of research. Through gTME, it is now possible to unlock complex phenotypes regulated by multiple genes which would be very unlikely to reach by the relatively inefficient, iterative gene-by-gene search strategies. The concept of gTME is generic and provides access points for diverse transciptome modifications broadly impacting phenotypes of higher organisms too, as further studies with yeast amply demonstrate.

On the basis of these studies, combinatorial methods are generally more powerful in obtaining a given cellular objective than systematic methods due to their ability to make broader perturbations. However, properly designed search strategies which make use of both systematic and combinatorial approaches may be the best route for optimizing phenotypes.

Thesis Supervisor: Gregory Stephanopoulos
Title: Bayer Professor of Chemical Engineering

Acknowledgments

First, I would like to thank my thesis advisor, Greg Stephanopoulos for his support and guidance throughout my PhD research. I would also like to express my appreciation for the members of my thesis committee: Boris Magasanik, Kristala Jones Prather, Pierre Rouviere, Daniel I.C. Wang, and K. Dane Wittrup. Through thesis meetings and informal discussions, these people have helped to provide me with invaluable advice and guidance.

Research is not cheap, and I am grateful to the DuPont-MIT alliance for providing financial support for the work as well as for providing numerous occasions to interact with scientists at DuPont.

Research is not something that one does for fun in solitude (at least most people), and I owe thanks to several people who have helped me throughout my thesis. First on this list is, appropriately, the first person I met while visiting MIT: Joel Moxley. Many thanks go to Joel for providing invaluable collaborations and discussions throughout our entire time at MIT. Thanks is also due to: Curt Fischer for our time working together especially with the promoter work; Keith Tyo for always being around to listen to ideas and for providing input; Kohei Miyaoku for helping out with the bioreactors and making the overnight experiments more enjoyable; and the remainder of the members of the Stephanopoulos lab for help throughout my thesis.

Finally, none of this work would have been possible without the love and support of my parents and sisters, and for that I am forever indebted.

Contents

Abstract *3*

Acknowledgments *5*

List of Figures *12*

List of Tables *16*

1. Introduction *17*

1.1 Motivation **17**

1.2 Objectives **19**

1.3 Approach **20**

1.4 Thesis Organization **21**

2. Metabolic Engineering Overview *23*

2.1 Evolution of the Metabolic Engineering Approach **24**

2.2 Systematic Approaches **25**

2.2.1 Data-driven approaches 26

2.2.2 Model-based approaches 27

2.3 Combinatorial Approaches **30**

2.3.1 Inverse Metabolic Engineering Paradigm 30

2.3.2 Tools for introducing genetic perturbations 31

3. Lycopene Bioproduction *33*

3.1 Isoprenoid pathway **36**

3.1.1 Non-mevalonate pathway 36

3.1.2 Mevalonate pathway 39

3.2 Carotenoid pathway 40

3.3 Summary 41

4. Systematic target identification 42

4.1 Stoichiometric Model 43

4.2 Identification of putative parameters 47

4.2.1 Growth rate and glucose uptake 47

4.2.2 Formate production 47

4.2.3 Oxygen uptake rate 51

4.2.4 Carbon source optimization 52

4.2.5 Summary 52

4.3 Single gene knockout targets 54

4.3.1 Single knockout genome scan 54

4.3.2 Linking of *gdhA* and NADPH 58

4.4 Multiple gene knockout targets 62

4.4.1 Multiple knockout identification 62

4.4.2 Sequential vs. simultaneous searches 65

4.5 Experimental validation of targets 68

4.5.1 Conclusions from results 70

4.5.2 Comparison to random perturbations 71

4.6 Summary 74

5. Mapping the metabolic landscape 77

5.1 Identifying combinatorial targets 78

5.2 Mapping the metabolic landscape 80

5.2.1 Creating a systematic-combinatorial metabolic landscape 80

5.2.2 Visualizing the metabolic landscape 83

5.3 Uncovering genetic interactions 87

5.3.1 Impact of combinatorial targets 87

5.3.2 Hierarchical clustering analysis 90

5.3.2.1 Stoichiometric targets have similar modes-of-action 90

5.3.2.2 Combinatorial targets decouple modes of action of stoichiometric targets 92

5.3.2.3 Clustering analysis highlights varied modes-of-action 95

5.3.3 Covariance analysis 97

5.3.4 Summary 100

5.4 Optimizing fermentation profiles 100

5.5 High cell density fermentations 103

5.5.1 Determination of optimal fermentation parameters 103

5.5.1.1 Agitation 103

5.5.1.2 pH control 105

5.5.2 High cell density fermentations 108

5.5.3 Carbon balances 113

5.5.4 Bioreactor summary 116

5.6 Summary 117

6. The lycopene gene knockout search network 121

6.1 Probing the metabolic landscape 122

6.2 Creating a search network diagram 129

6.3 Understanding network complexity 132

6.3.1 Gateway nodes 132

6.3.2 Model accessible nodes 134

6.3.3 Mostly model inaccessible nodes 135

6.3.4 Model inaccessible nodes 136

6.4 Further characterization of strains 136

6.4.1 *yliE* investigation 139

6.5 Summary 142

7. Promoter Engineering 144

7.1 Motivation 144

7.2 Background 145

7.3 Implementation 147

7.3.1 Multi-faceted characterization 151

7.3.2 Promoter strength metric 154

7.4 Applications 155

7.4.1 Growth yield and *ppc* activity 155

7.4.2 Lycopene yield and *dxs* activity 157

7.4.3 Lycopene yield and *ppc* activity 161

7.5 Summary 163

8. global Transcription Machinery Engineering (gTME) 166

8.1 Motivation 166

8.2 Background 167

8.3 Implementation 169

8.4 *E. coli* Applications 171

8.4.1 Ethanol tolerance 171

8.4.1.1 Transcriptional analysis 177

8.4.2 Lycopene Production 186

8.4.3 Multiple tolerances 191

8.4.4 Acetate tolerance 195

8.4.5 pHBA tolerance 199

8.4.6 Hexane tolerance 201

8.4.7 E. coli summary 203

8.5 Yeast applications 203

8.5.1 LiCl tolerance 205

8.5.2 High glucose tolerance 207

8.5.3 Ethanol and glucose multiple tolerance 209

8.6 Summary 211

9. Conclusions and recommendations 214

9.1 Summary 214

9.2 Conclusions 217

9.3 Recommendations for future work 220

10. Materials and methods 223

10.1 Commonly used techniques 223

10.1.1 Flux balance analysis calculations 223

10.1.1 Lycopene Assay 224

10.1.1 Transposon library generation and screening 224

10.1.1 Gene knockout construction and verification 225

10.2 Systematic gene knockouts 225

10.2.1 Strains and media 226

10.2.2 Primers for gene knockouts 226

10.3 Metabolic landscape 228

10.3.1 Strains and media 228

10.3.2 Hierarchical Clustering Routines 229

10.4 High cell density fermentations 229

10.4.1 Fermentation conditions 229

10.4.2 Organic and amino acid measurements 230

10.5 Probing the metabolic landscape 231

10.5.1 Strains and media 231

10.6 Promoter Engineering 232

10.6.1 Strains and media 232

10.6.2 Library construction 233

10.6.3 Library characterization 233

10.6.3.1 Initial characterization 233

10.6.3.2 Promoter strength metric 234

10.6.3.3 Transcriptional analysis 235

10.6.3.4 Chloramphenicol resistance 235

10.6.4 Promoter delivery 235

10.6.5 List of primers 236

10.7 global Transcription Machinery Engineering 237

10.7.1 Strains and media 237

10.7.2 Library construction 238

10.7.3 Sequence analysis 239

10.7.4 Transcriptional analysis 239

10.7.5 Phenotype selection 240

10.7.6 Yeast Examples 241

11. References 243

List of Figures

3.1 Structure of lycopene 33

3.2 Lycopene production pathway 35

4.1 Impact of growth rate and glucose uptake rate on lycopene stoichiometric yield 49

4.2 Impact of formate production on lycopene stoichiometric yield 50

4.3 Impact of oxygen on lycopene stoichiometric yield 51

4.4 Carbon source optimization 53

4.5 *E. coli* genome scan for single gene knockout targets 56

4.6 *in silico* NADPH utilization in a wild-type strain 59

4.7 *in silico* NADPH utilization in a *gdhA* knockout strain 60

4.8 NADPH production rates 61

4.9 Identification of sequential, multiple gene knockout targets 64

4.10 Simultaneous approach to multiple target identification 67

4.11 Comparison of selected mutants to random libraries of knockouts 73

5.1 Comparison of combinatorial targets to parental strain 79

5.2 Systematic and combinatorial gene knockout target identification 82

5.3 Visualization of the metabolic landscape at 15 hours 84

5.4 Visualization of the metabolic landscape at 24 hours 85

5.5 Visualization of the metabolic landscape—maximum production 86

5.6 Impact of combinatorial genotypes on systematic backgrounds 89

5.7 Clustering analysis of time course data for systematic targets 91

5.8 Clustering analysis of time course data for combinatorial targets 93

5.9 Clustering analysis of time course data for systematic targets with *hnr* knockouts .. 94

5.10 Clustering analysis and bubble plots 96

5.11 Covariance analysis of systematic targets 98

5.12 Covariance analysis of combinatorial targets 99

5.13 Behavior of selected strains in optimized culturing conditions 102

5.14 Fermentation-based investigation of oxygen level on lycopene 104

5.15 Impact of double-sided pH control on lycopene production 106

5.16 Impact of single-sided pH control on lycopene production 107

5.17 Volumetric lycopene production in the reactors 110

5.18 Specific lycopene production (ppm) in the reactors 111

5.19 Overall carbon yield balances for the fermentors 114

5.20 Marginal carbon yield balances for fermentors 115

6.1 Original search network diagram for the metabolic landscape 130

6.2 Complete search network diagram for the metabolic landscape 131

6.3 Lycopene production of selected strains at 15 and 24 hours in 1xM9 138

6.4 Comparison of Δ*hnr* Δ*yliE* to previous maxima at 15 and 24 hours in 2xM9 140

6.5 Impact of Δ*yliE* in other genotypes 141

7.1 Generation of the functional promoter library 149

7.2 Flow cytometry analysis of the functional promoter library 150

7.3 Comprehensive characterization of the promoter library 153

7.4 Growth yield and *ppc* activity 156

7.5 Lycopene yield and *dxs* activity in wild-type K12 159

7.6 Lycopene yield and *dxs* activity in a pre-engineered strain 160

7.7 Lycopene yield and *ppc* activity 162

8.1 Basic methodology of global transcription machinery engineering 170

8.2 Overall improvement of ethanol tolerance using gTME 172

8.3 Sequence analysis of ethanol sigma factor mutants 174

8.4 Growth curves for ethanol-tolerant sigma factor mutants 176

8.5 Transcriptional analysis of general ethanol stress 179

8.6 Transcriptional analysis of an ethanol sigma factor mutant 180

8.7 Transcriptional analysis of an ethanol sigma factor mutant in response to ethanol. 181

8.8 Patterns in the transcriptional profiles in response to ethanol 185

8.9 Sequences for lycopene sigma factor mutants 187

8.10 Application of gTME to a metabolite production phenotype 188

8.11 Genotype specificity of identified sigma factor mutants 190

8.12 Eliciting multiple, simultaneous phenotypes using gTME 193

8.13 Sequence analysis of multiple, simultaneous phenotypes using gTME 194

8.14 Growth analysis of acetate mutants 197

8.15 Sequence analysis of acetate sigma factor mutants 198

8.16 Growth analysis of pHBA sigma factor mutants 200

8.17 Sequence analysis of pHBA sigma factor mutants 200

8.18 Growth analysis of hexane sigma factor mutants 202

8.19 Sequence analysis of hexane sigma factor mutants 202

8.20 Growth analysis of LiCl gTME mutants in yeast 206

8.21 Sequence analysis of LiCl gTME mutants in yeast 207

8.22 Growth analysis of glucose gTME mutants in yeast 208

8.23 Sequence analysis of glucose gTME mutants in yeast 209

8.24 Growth analysis of ethanol-glucose gTME mutants in yeast 210

8.25 Sequence analysis of ethanol-glucose gTME mutants in yeast 211

List of Tables

3.1 Non-mevalonate pathway for isoprenoid biosynthesis 38

3.2 Mevalonate pathway for isoprenoid biosynthesis 39

3.3 Carotenoid pathway 41

4.1 Modifications of the *i*JE660a model 45

4.2 Metabolite abbreviations for the modified *i*JE660a model 46

4.3 Experimental results of single and multiple gene knockouts 69

5.1 Growth and lycopene phenotypes of strains in fed-batch reactor 112

6.1 Identified gene knockouts in the parental strain background 123

6.2 Identified gene knockouts in the Δ*gdhA* Δ*aceE* background 123

6.3 Identified gene knockouts in the Δ*gdhA* Δ*aceE* Δ*fdhF* background 124

6.4 Identified gene knockouts in the Δ*gdhA* background 124

6.5 Identified gene knockouts in the Δ*yjfP* background 125

6.6 Identified gene knockouts in the Δ*hnr* background 125

6.7 Identified gene knockouts in the Δ_{P}*yjiD* background 126

6.8 Identified gene knockouts in the Δ*gdhA* Δ*aceE* Δ_{P}*yjiD* background 126

6.9 Fold improvement in lycopene production by identified gene knockouts 127

8.1 Improvement of ethanol tolerance through engineered sigma factors 175

8.2 Change in expression of ethanol response genes 182

8.3 Change in expression of sigma factor mutant-induced genes 182

8.4 Change in expression of new ethanol response 183

10.1 Primer Designs for Gene Knockout Constructs 227

Chapter 1

Introduction

1.1 Motivation

The improvement of cellular properties using modern genetic tools is a central goal of metabolic engineering (Stephanopoulos, 2002). Advances in molecular biology and genetic engineering empower metabolic engineers with the increasing ability to create any desired cellular modification. These new tools complement the global focus to target identification which has always been a strength of the metabolic engineering paradigm. Embedded in these concepts is the understanding that cellular phenotype reflects *global* intracellular conditions, not *individual* gene states. Beyond individual metabolite pathways, cellular phenotype is a manifestation of global gene expression levels, metabolic demand, resource availability, and cellular stresses. Above all, metabolic function is constrained by the stoichiometry and individual reaction kinetics of the reaction network. This fundamental understanding has been at the heart of metabolic engineering since its conception over a decade and a half ago. However, since its conception, a particularly useful, additional tool has become available for metabolic

engineers. The recent availability of whole-genome sequences can greatly assist and alter the process of conducting such system-wide analyses.

The advent of genome sequencing has greatly expedited the discovery process. However, genome sequences and catalogues of bioreaction networks only provide a list of parts to be used in this endeavor. Beyond these, complexities and nonlinearities in the interactions of metabolic pathways and regulatory networks confound the process of cellular and metabolic engineering. To accomplish these tasks in an efficient and comprehensive manner, a diverse set of molecular biology tools must accompany, and at times supplement, systematic analysis of pathways. These tools and methodologies must be both broad in effect (since different genes require different levels of modification) and in scope (since each pathway has a unique set of regulatory bounds). When collectively used, these advances in molecular biology and genetic engineering enable the realization of whole-cell engineering. Consequently, the development of methods to identify key genetic targets and subsequently, the ability to make broad modifications are required to accomplish the broad goals of whole-cell engineering.

Once created, these tools can be linked with high-throughput screening to help unlock latent cellular phenotypes and ultimately, lead to the understanding of genotype-phenotype relationships. However, a set of tools will only be as effective as the context in which they are used. As such, efficient phenotype optimization necessitates a robust, defined *search strategy to identify genetic targets* requiring modification. By exploring and probing the metabolic landscapes created by the underlying structure of genotype-phenotype interaction, lessons may be gained which can help guide future cellular and metabolic engineering programs. Furthermore, it is not clear how one should approach a

given problem in metabolic and cellular engineering. A complement of tools, both systematic and combinatorial, is available, yet it is unclear how these tools should be used to bring about the most substantive changes to a cellular system. In this light, this thesis addresses the issue of the development of both strategies and tools for the identification of genetic targets for the engineering of microorganisms.

1.2 Objectives

To accomplish the goals set out for this thesis, two major objectives were proposed:

- Evaluate the applicability of systematic approaches to the case study of increasing lycopene yield
    - Evaluation of stoichiometric models as a means of navigating the metabolic landscape
    - Evaluation of combinatorial tools for the further optimization of metabolic phenotypes
    - Evaluation of search strategies to elucidate the topology of the gene knockout search network
- Develop and evaluate combinatorial tools which modulate gene expression and regulatory networks
    - Development of a tool for the optimization of gene expression level, applicable for chromosomal-level modifications
    - Development of a tool for the combinatorial alteration of regulatory networks and simultaneous alterations of multiple genes

1.3 Approach

To address the problem of understanding and improving methods for the identification of gene targets, we have focused on the study of recombinant lycopene production pathway in *Escherichia coli*. In particular, this study focused on the identification of gene knockout targets, however, it is emphasized early that the analysis and results obtained are not limited to this mode of perturbation. As such, the strategies and lessons may be easily applied to other systems such as gene overexpression or other modes of perturbations as described later. The approach consists of using a global, stoichiometric model to identify single and multiple gene knockout targets. This target selection is then complemented through the use of transposon mutagenesis to identify a disparate set of gene knockout targets. These two sets are then combined to gain an understanding of the metabolic landscape. Finally, this landscape is analyzed at various important nodes through additional transposon mutagenesis searches. The resulting analysis will present a picture and understanding of the resulting gene knockout search network. It will be demonstrated that despite the high degree of complexity in these systems, certain key nodes are universal and thus serve as platforms for strain improvement.

Two novel tools for metabolic engineering will also be discussed and demonstrated through a number of examples. The first, a tool for optimizing gene expression, termed Promoter Engineering will be addressed. After demonstrating the development of a fully-characterized, wide dynamic range of constitutive promoter strengths, a library will be constructed. The utility of this library will be demonstrated

through the optimization of two gene expression levels, *dxs* and *ppc* for the increase of lycopene yield and both cell yield and lycopene yield respectively. The applicability of this tool to other host systems will be discussed.

Finally, the tool of global Transcription Machinery Engineering (gTME) will be demonstrated as a novel tool for the introduction of multiple, simultaneous modifications to gene expression. The methodology for this approach will be presented followed by a series of examples of improvement of phenotypes in *E. coli*. Furthermore, the applicability of this tool to other host systems will be demonstrated through examples of phenotype improvement in *Saccharomyces cerevisiae*.

In general, the approach to accomplishing the broad goals outlined in Section 1.3 is two-pronged: (1) Investigation of systematic approaches to metabolic engineering problems and (2) Development of tools which can aid in metabolic engineering efforts. As such, the research presented in this thesis will reflect these two broad areas.

1.4 Thesis Organization

The metabolic engineering paradigms of systematic and combinatorial approaches are presented in Chapter 2 as an overview of the current portfolio of successful attempts at engineering cellular systems. A major model system utilized for this study was the recombinant bioproduction of lycopene in *Escherichia coli*. An overview of the lycopene production pathway and prior attempts of engineering this model system is presented in Chapter 3. Chapters 4 through 6 address the application of various techniques and search strategies for the identification of gene targets and subsequent engineering of *E. coli* for

the production of lycopene. Chapter 7 presents the development of Promoter Engineering, a tool for the optimization of gene expression. Chapter 8 will present a tool termed global Transcription Machinery Engineering (gTME) which provides for the simultaneous, multiple modification of the transcriptome. Finally, this thesis concludes with a discussion of the impact of these results (Chapter 9) with respect to the metabolic engineering paradigms presented in Chapter 2 as well as recommendations for further studies. A comprehensive Materials and Methods section will be included in Chapter 10 to cover all experiments in this thesis.

Chapter 2

Metabolic Engineering Overview

Metabolic engineering is a young field, nearly fifteen years old. During this period, it has developed a well-defined methodology and a focused research portfolio of rich intellectual content and particular relevance to biotechnology and biological engineering. New and diverse opportunities for metabolic engineering emerge quickly in this post-genomics era. These opportunities provide a challenge to the metabolic engineering paradigm. In particular, the scope of problems posed to the field is rapidly increasing in complexity. Although the focus (e.g. improving cells) and a central component (e.g. assessing cell physiology) of metabolic engineering remain the same, new tools are required to take advantage of the opportunities arising from the availability of whole-genome sequence information. This chapter will review the evolution of the metabolic engineering approach, and in particular, will highlight various systematic and combinatorial approaches previously used to optimize strains and identify genetic targets for a given phenotype. The purpose of this section is to briefly provide a context for the results presented in chapters 4-9.

2.1 Evolution of the Metabolic Engineering Approach

The current portfolio of advances in metabolic engineering is large for such a young field of study. Concepts and methodologies have been applied to extending cellular substrate ranges (Becker & Boles, 2003; Ostergaard et al., 2000; Prieto, Diaz, & Garcia, 1996), increasing product yields (Koffas, Jung, & Stephanopoulos, 2003), and diversifying product ranges (Cameron et al., 1998; Farmer & Liao, 2000; Watanabe et al., 2003). In addition to solely dealing with substrate and product diversification, advances have been made in balancing reduction potentials within a cell (Berrios-Rivera, Bennett, & San, 2002) and improving the ability of cells to thrive in non-traditional environments such as hypoxic (Khosla & Bailey, 1988) and toxic (J-Y Lee, Roh, & Kim, 1994) conditions. Furthermore, although frequently focused on bioprocessing applications, the broad applicability of metabolic engineering concepts has impacted research in the fields of biocatalysis (Stafford et al., 2002b) and medicine. Studying the metabolism of organs and cells has aided in the identification of genetic targets for disease therapy and the understanding of metabolic function for some disease states (Kyongbum Lee et al., 2003; Yarmush & Banta, 2003).

These examples illustrate the central focus of metabolic engineering. More specifically, these studies attempt to (1) identify genetic targets, (2) rigorously quantify metabolic phenotype, and (3) understand kinetic control in metabolic networks. Genetic targets may be identified systematically through determining the rate controlling step in a reaction or combinatorially through high-throughput screening. Once the identified genetic perturbations have been performed, high-throughput metabolic profiling tools aid in fully quantifying the resulting metabolic phenotype.

2.2 Systematic Approaches

Genomic sequences have facilitated the construction of cellular metabolism models enabling systematic approaches to gene target identification. Given the absence of extensive knowledge about the kinetics of molecular interactions, the dissection and optimization of metabolic pathways is an outstanding issue of central importance to metabolic engineering. These models are most often not of *kinetic* nature due to limited rate and regulatory data. As a result, stoichiometric models have been formulated where the pathways fluxes (reaction rates) are determined such as to optimize a pre-selected objective function (Kauffman, Prakash, & Edwards, 2003). Models and results based on bioreaction network stoichiometry provide a direction for modulating metabolism. To this end, putative parameters and interacting pathways may be extracted.

However, models based solely on reaction stoichiometry neglect entire portions of the genome responsible for regulation and control. Current models hold the promise of predicting the metabolic function of whole cells, especially when used in conjunction with other protein and small metabolite data. The integration and understanding of the cellular components gives rise to information about system behavior. Genomic sequencing information provides a catalogue of an organism's capacity and metabolic capability. Physiology and phenotype rely on the interactions and concentrations of all of these components which highlight the importance of integrating multiple dimensions of molecular interactions in order to predict global, system response. Nevertheless, these models have proved invaluable in probing cellular systems for putative parameters and

focal points of cellular metabolism. These systematic approaches may be viewed as being either data-driven or purely model-based.

2.2.1 Data-driven approaches

Comprehensive metabolic profiling requires measuring metabolite levels and reaction fluxes, typically through the use of a gas chromatography-mass spectrometry (GC-MS) unit to detect metabolite levels. Used in conjunction with isotopic labeled substrates, GC-MS spectra provide insight into the distribution of the labeled substrate through the various pathways of the bioreaction network. Further advances in high-throughput metabolite and isotopic measurements (Soga et al., 2003) will continue to advance our ability to probe the underlying factors influencing cellular phenotype, in particular, abilities to measure metabolite pools and pathway fluxes. Once these variables are adequately assessed, kinetic control may be elucidated through further genetic perturbations and measurement of the metabolic response. These concepts collectively embodied in Metabolic Control Analysis (MCA) (Stephanopoulos, Aristidou, & Nielsen, 1998) can help elucidate the link between genotype and phenotype, while at the same time, identify future gene targets. The major principle behind MCA resides in creating fluctuations within the cell for various enzyme levels and measuring the impact on a certain factor, like a production rate (Stafford et al., 2002a; Stephanopoulos, 1999). As an example, flux data can be used to identify various nodes and distributions of carbon flux in a cellular system to identify required gene knockouts or overexpressions (Colon et al., 1995).

Other forms of high-throughput data have been used to extract putative genetic targets. In one such application, association discovery was employed for the evaluation of a library of unsequenced fungal strains of *Aspergillus terreus* for their ability to over-produce the antibiotic lovastatin (Askenazi et al., 2003). Through the use of gene overexpressions, a large diversity of strains was generated with respect to the production profiles of lovastatin and (+)-Geodin and the strains were characterized by metabolite and transcriptional profiling. These measurements generated a wealth of biological data, from which Askenazi *et al.* were able to extract key putative parameters and genes by performing a statistical association analysis. Ultimately, this data-based approach can make use of high-throughput data collection to predict the next perturbation necessary for a given cellular state. However, these techniques often require a significant amount of experimental work and the results are not ordinarily extensible to other genotypes.

2.2.2 Model-based approaches

As a complementary approach to traditional laboratory experiments, metabolic simulations are becoming a useful tool for probing cellular function. Current computational methods simulating metabolism (*in silico* predictions) attempt to probe cellular function by simulating the bioreaction network. Ultimately, a comprehensive kinetic model of metabolism could aid in the identification of genetic targets and putative parameters influencing phenotype. Models can range from strictly reflecting stoichiometry (Edwards & Palsson, 2000) to detailed enzymatic kinetics of an entire pathway (Wiechert, 2002). In these efforts, metabolic engineering has borrowed heavily

from the framework of traditional chemical reaction engineering. Most advances in chemical reaction engineering require a model reflective of the system dynamics (such as a rate expression). While these tools could be easily applied to cellular systems, the limitations of current cellular models severely limit the amount of information we can extract from the models. Despite limited data about the intracellular conditions and kinetic parameters, many dynamic models have been assembled for gene expression modeling and for several, well-characterized systems (Chen, 1997; Fell, 1998; Niederberger et al., 1992; Stafford et al., 2002a). Borrowing from the simplifying assumptions of chemical reaction engineering, our lack of understanding is often masked by concepts such as "rate-limiting steps" and "functions of genes" (J. Bailey, 1999). However, these reduced models do not always accurately model all *in vivo* cellular response and require improvement by further experiments.

A great deal of emphasis has been recently placed on using stoichiometric models for the determination of putative parameters and gene knockout targets for many cellular systems including *Escherichia coli*, *Saccharomyces cerevisiae*, *Synechocystis sp.* (Edwards, Ibarra, & Palsson, 2001; Edwards & Palsson, 2000; Famili et al., 2003; Forster et al., 2003; Shastri & Morgan, 2005). These methods of flux balance analysis revolve around the basic principle of applying the steady state solution to the dynamic metabolite balance. When the steady state assumption is invoked, the transient metabolite balance (a differential equation) assumes the form of a linear matrix expression:

$$S \bullet v = b \qquad \text{(Equation 2.1)}$$

where S is the stoichiometry matrix, v is a vector of fluxes, and b is a vector of transport rates into the cell. However, these systems are extremely underdetermined with the

number of fluxes on the order of a thousand and number of metabolites in the S matrix on the order of hundreds. To solve this underdetermined system, it is necessary to create an objective function and a typical approach is to use linear programming to determine the fluxes given a series of flux constraints. Often times, maximization of biomass production serves as the exclusive objective function used to solve the matrix equation (Edwards, Ibarra, & Palsson, 2001; Kauffman, Prakash, & Edwards, 2003).

However, for systems in which genetic perturbations (knockouts or over-expressions) are introduced, the resulting phenotype is often suboptimal. To calculate the flux profile in suboptimal systems, a minimization of metabolic adjustment (MOMA) calculation serves as an additional constraint in which the resulting flux profile is intermediate between the wild-type optimal and mutant optimal and requires a quadratic programming to solve (Segre, Vitkup, & Church, 2002). However, these sets of constraints and objective functions is not exhaustive as several attempts have been made to include a bi-level optimization to provide for dual optimization of cellular and bioengineering objectives (Burgard, Pharkya, & Maranas, 2003). Furthermore, several attempts have been made in attempts to further restrict the resulting fluxes including the addition of thermodynamic constraints to impose restrictions on this underdetermined system (Beard, Liang, & Qian, 2002). Newer versions of stoichiometric models have made attempts to include regulation and further refinement of stoichiometric reactions including specificity of redox pairs (Reed et al., 2003). Finally, recent advances have attempted to improve the methods for calculating the suboptimal fluxes resulting in a gene knockout by limiting the number of fluxes changing after genetic perturbations (Shlomi, Berkman, & Ruppin, 2005). Each of these methods aim to solve the same

problem of obtaining cellular properties and putative genetic targets using only knowledge of genome sequences and biochemical reactions available in a cell. The application and demonstration of these models, however, has received little attention experimentally.

2.3 Combinatorial Approaches

Typically, gene deletions and amplifications serve as effective tools for genetic modification. The intrinsic link between cellular genotype and phenotype may be extracted by studying the response of cells to these systematic changes. In recognition of the importance of these changes in probing the genotype-phenotype relationship, a diverse set of tools have emerged to create these specified genetic modifications. Molecular biology advances have provided the ability to perform these modifications at will. In addition to gene-specific tools, a number of combinatorial tools have also been created which, when combined with high throughput screening, allow for randomized gene expression levels (including deletions) and genomic library complementation.

2.3.1 Inverse Metabolic Engineering Paradigm

The identification of genetic targets through a systematic approach is often a very difficult problem due to the lack of metabolic models apt at capturing both reaction kinetics and genetic regulation. An alternative method for introducing cellular perturbations to identify targets, termed inverse metabolic engineering (J. E. Bailey et al., 2002), uses introduced perturbations linked together with high throughput screening to

ultimately identify genetic targets. This methodology uses the approach of screening for a desirable phenotype using a perturbation library and tracing genetic modifications responsible for the cellular response. The main objective of the inverse approach is to identify targets which, following modification, will elicit a desired phenotype rather than randomly evolving a high product titer strain. Moreover, recent advances in genomics technologies allow for this process to proceed via high throughput screening methods (Badarinarayana et al., 2001; Gill et al., 2002). One good example is parallel gene trait mapping (PGTM), which exploits DNA microarray as a tool for the high-throughput identification of genes conferring a particular phenotype (Gill, 2003). Furthermore, tools such as flow cytometry and microfluidic devices allow for high-throughput screening and selection of mutants with improved cellular properties. Beyond these tools, perhaps one of the greatest experimental advances has been genome sequencing. The ready availability of sequence data extends the impact of metabolic engineering. The development of tools for diverse genetic perturbations is necessary to uncover critical gene targets.

2.3.2 Tools for introducing genetic perturbations

The introduction of tools for genetic perturbations includes the introduction of randomized gene knockouts (Badarinarayana et al., 2001) and overexpressions through shotgun genomic libraries (Kang et al., 2005). Most tools for genetic perturbations have benefited from genome sequencing efforts. The ability to obtain sequence information influences the experimental tools used to understand and modify cells. As an example, transposons, which allow for randomized genomic knockouts, have become a tool for

studying the general relationship between genotype and phenotype not only in microorganisms, but also in higher eukaryotes such as mice (Hayes, 2003). Screening libraries of transposon knockout strains and subsequent DNA sequencing can identify unknown genes influencing a particular phenotype (Hemmi et al., 1998). The direct ability to perform these randomized knockouts and identify their location by sequencing has undoubtedly increased the throughput of such experiments and has lead to an incentive to search for the minimal genome (Hutchison et al., 1999), which may be of use for bioprocessing applications. Furthermore, bioinformatics tools can search sequence data to identify particular elements within the non-coding regions which are important for cellular function. As an example, elements such as microRNAs and interfering RNA (RNAi) may be predicted from bulk sequence data (Lewis et al., 2003). Once identified, these elements have the potential to be powerful molecular biology tools for gene silencing. Finally, DNA sequencing facilitates efforts in directed evolution and mutagenesis. The generation of beneficial mutants and subsequent sequencing can help in reducing the search space in future studies and lead to a further understanding of sequence-function relationships and patterns. Ultimately, these tools may be used to create perturbations which can elicit phenotypes of interest. However, it is unclear how to efficiently and effectively utilize these tools. Furthermore, very few tools allow for the fine-tuning of genetic control and fewer tools address the important consideration of modifying multiple genetic targets simultaneously.

Chapter 3

Lycopene Bioproduction

Lycopene is a hydrocarbon molecule that may be classified as a carotenoid. Carotenoids are molecular members of the isoprenoid family of compounds found within all cells. Isoprenoids larger than 5 carbon units are formed through the head-to-tail condensation of multiple isoprene units to create the desired length. Lycopene ($C_{40}H_{56}$) is a red pigment carotenoid possessing a characteristic conjugated, aliphatic hydrocarbon chain shown in **Figure 3.1**.

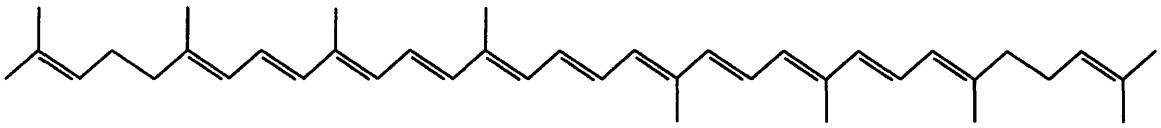

Figure 3.1: Structure of lycopene. Lycopene is a 40 carbon conjugated hydrocarbon which serves as a gateway molecule into other cyclic carotenoids.

In addition to optical properties, the conjugated structure of lycopene and similar carotenoids has been implicated in mechanisms related to photodynamic action protection and singlet oxygen quenching (Sandmann, 2002). These molecules have received a

significant amount of attention in recent years as a result of their antioxidant, UV protecting, and, natural food colorant properties (P. C. Lee & Schmidt-Dannert, 2002). Furthermore, large families of carotenoids containing cyclical structures (including β-carotene) require lycopene as a precursor.

Naturally, carotenoids are produced in plants and fungal systems (Cunningham & Gantt, 1998). The recent elucidation of the various metabolic routes and required enzymes for carotenoid production has allowed for the possibility of high-level production of diversified carotenoid molecules (both naturally occurring and synthetic) in recombinant hosts such as *Escherichia coli*. Roughly, the metabolism of these molecules in a host cell may be divided into three areas: (1) the isoprenoid pathway, (2) the carotenoid pathway, and (3) remainder of cellular metabolism, which supplies precursors and cofactors required for the production of this expensive, secondary metabolite. **Figure 3.2** summarizes the lycopene production pathway for the non-mevalonate route including (in various detail) all three contributing pathways. Furthermore, the overall stoichiometry is included which indicates that for every molecule of lycopene produced, 16 NADPH reducing equivalents are required in addition to 8 CTPs and 8 ATPs. This stoichiometry illustrates the high energetic and redox requirement required to produce lycopene.

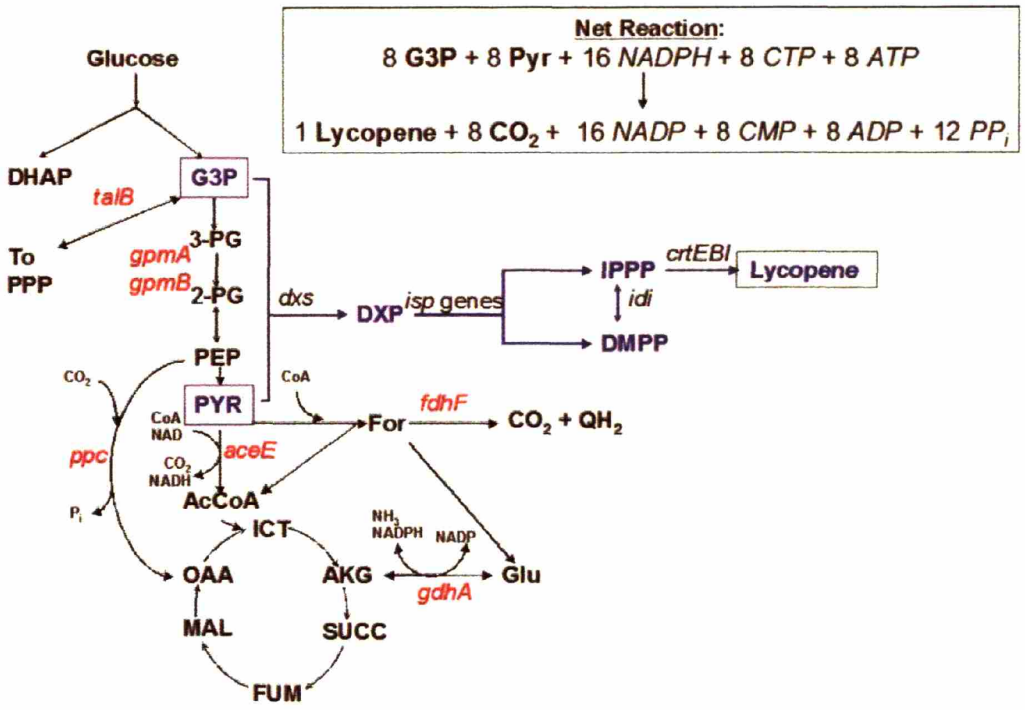

Figure 3.2: Lycopene production pathway. Lycopene synthesis begins with the condensation of the key glycolytic intermediates, glyceraldehyde 3-P (G3P) and pyruvate (PYR) and continues in a nearly linear pathway. The genes encoding for *idi* and *dxs* are typical targets for lycopene over-expression along this pathway. The overall stoichiometry of this reaction is included, which highlights the high energetic and redox requirements for biosynthesis.

3.1 Isoprenoid pathway

Engineering strains for the overproduction of isoprenoid-based molecules is of significant interest due to the diversified base of products accessible through this molecular gateway. Branching from the five-carbon precursor unit of isopentenyl-pyrophosphate (IPPP), it is possible to create carotenoids, quinones, and even precursors for desirable pharmaceuticals such as Taxol and Artemesinin (Huang et al., 2001; Martin et al., 2003). Lycopene is a major precursor to downstream, modified carotenoids (Sandmann, 2002). Lycopene production in *E. coli* requires the heterologous expression of the *crtEBI* genes to encode the polymerization of IPPP to the 40 carbon molecule of lycopene. However, two possible routes exist for the synthesis of IPPP in nature, termed the mevalonate and the non-mevalonate pathway. For the studies presented in this thesis, we investigated the issues of gene target identification in the context of heterologous lycopene production in *E. coli* using the non-mevalonate pathway (Adam et al., 2002).

3.1.1 Non-mevalonate pathway

Isoprenoid production using the non-mevalonate pathway in *E. coli* utilizes glycolytic intermediates to form precursor monomers which subsequently undergo polymerization to form the 40 carbon biopolymer. Through the use of this pathway, two precursor molecules, glyceraldehyde-3-phosphate and pyruvate are used to form the predominant isoprenoid unit, isopentenyl pyrophosphate (IPPP). **Table 3.1** summarizes

the various reaction steps required to convert the precursors into IPPP. In general, a seven reaction series is required for the conversion of glyceraldehyde-3-phosphate and pyruvate to IPPP (Adam et al., 2002; Hecht et al., 2001). Initial attempts for improving carotenoid production in *E. coli* targeted the expression of genes coding for enzymes for this non-mevalonate pathway (Farmer & Liao, 2000, 2001; P. C. Lee & Schmidt-Dannert, 2002). Even with the over-expression of *dxs* and *idi* genes (Kajiwara et al., 1997; Seon-Won Kim & Keasling, 2001; Mathews & Wurtzel, 2000), cellular production and accumulation of carotenoids were limited by regulatory networks and precursor supply (Farmer & Liao, 2000, 2001; Jones, Kim, & Keasling, 2000; P. C. Lee & Schmidt-Dannert, 2002; Wang, Oh, & Liao, 1999).

Pathway	Gene	Reaction
Isoprenyl-pyrophosphate synthesis pathway	*dxs*	G3P + Pyruvate $\rightarrow$ DXP + CO_2
	ispC	DXP + NADPH $\leftrightarrow$ MEP + NADP
	ispD	MEP + CTP $\rightarrow$ CDPME + PPI
	ispE	CDPME + ATP $\rightarrow$ CDPMEPP + ADP
	ispF	CDPMEPP $\rightarrow$ MECPP + CMP
	ispG	MECPP $\leftrightarrow$ HMBPP
	ispH	HMBPP $\rightarrow$ 0.5 IPPP + 0.5 DMPP
Isoprenyl pyrophosphate isomerase	*idi*	IPPP $\leftrightarrow$ DMPP

Molecule Abbreviation	Molecule Name
CDPME	4-diphophocytidyl-2-C-methyl-d-erythritol
CDPMEPP	4-diphosphocytidyl-2-C-methyl-2-phosphate-d-erythritol
DMPP	Dimethylallyl pyrophosphate
DXP	1-deoxy-d-xylulose-5-phosphate
G3P	Glyceraldehyde 3-phosphate
GPP	trans Geranyl pyrophosphate
HMBPP	1-hydroxy-2-methyl-2(E)-butenyl-4-diphsophate
IPPP	Isopentyl pyrophosphate
MECPP	2-methyl-d-erythritol-2,4-cyclodiphosphate
MEP	Polyol 2-C-methyl-d-erythritol-4-phosphate
PPI	Pyrophosphate

Table 3.1: Non-mevalonate pathway for isoprenoid biosynthesis. The non-mevalonate pathway includes a seven step reaction pathway from the precursors of glyceraldehyde-3-phosphate and pyruvate to the formation of IPPP. Further IPPP may be generated from DMPP through the isomerization reaction catalyzed by *idi*.

3.1.2 Mevalonate pathway

An alternative pathway for the production of isoprenoids found predominately in many eukaryotic systems is the mevalonate pathway, named for its metabolite intermediate. However, several laboratories have engineered the heterologous mevalonate pathway in *E. coli* (Campos et al., 2001; Martin et al., 2003). In this pathway, acetyl-coA serves as the sole precursor for the production of IPPP. A summary of the reactions for this pathway are provided in **Table 3.2**.

Pathway	Gene	Reaction
Mevalonate pathway to IPPP	*atoB*	2 A-CoA → AA-CoA + CoA
	hmgS	AA-CoA + A-CoA + H_2O → 3-HMG-CoA + CoA
	hmgR	3-HMG-CoA + 2 NADPH → Mev + 2 NADP + CoA
	Erg12	Mev + ATP → Mev-5P + ADP
	Erg8	Mev-5P + ATP → Mev-5PP + ADP
	Mvd1	Mev-5PP + ATP → IPPP + ADP + CO_2 + P_i
Isoprenyl pyrophosphate isomerase	*idi*	IPPP ↔ DMPP

Molecule Abbreviation	Molecule Name
A-CoA	Acetyl-CoA
AA-CoA	Acetoacetyl-CoA
3-HMG-CoA	3-hydroxy-3-methyl-glutaryl-coA
Mev	Mevalonate
Mev-5P	Mevalonate-5-phosphate
Mev-5PP	Mevalonate-5-diphosphate
IPPP	Isopentyl pyrophosphate

Table 3.2: Mevalonate pathway for isoprenoid biosynthesis. The mevalonate pathway

3.2 Carotenoid pathway

Regardless of the upstream pathway used, the precursor isoprene units of IPPP and the isomer version DMAP serve as the monomeric units for the production of carotenoids (as well as endogenously for quinones). The heterologous expression of a three enzyme complex results in the expression of the enzymes required for the polymerization to the C40 carotenoid molecule (Cunningham FX Jr, 1994; Umeno, Tobias, & Arnold, 2002). These three reactions are described in **Table 3.3**.

This portion of the isoprenoid pathway and downstream reactions have received great attention recently in an effort to create a diverse library of carotenoids (Sandmann, 2002; Sandmann et al., 1999; Schmidt-Dannert, Umeno, & Arnold, 2000; Umeno, Tobias, & Arnold, 2005). Through the discovery of novel enzymes with altered substrate specificity and through directed evolution and selection, novel carotenoids have been synthesized. These new molecules however, still require engineered cellular systems to create the necessary precursor molecules of IPPP and to provide the energy and redox cofactors which are required.

Pathway	Gene	Reaction
Lycopene Pathway	*crtE*	IPPP + FPP → GGPP + PPI
	crtB	2 GGPP → PHYTO + PPI
	crtI	PHYTO + 8 NADP → LYCO + 8 NADPH

Molecule Abbreviation	Molecule Name
FPP	trans, trans Farnesyl pyrophosphate
GGPP	Geranylgeranyl PP
IPPP	Isopentyl pyrophosphate
LYCO	Lycopene
PHYTO	Phytoene
PPI	Pyrophosphate

Table 3.3: Carotenoid pathway. The carotenoid pathway

3.3 Summary

Two pathways are available for the synthesis of precursors for the production of isoprenoid-based molecules in *E. coli*. These isoprenoid-based molecules are as diverse as cancer drugs, high valued coenzymes (such as coenzyme Q) and carotenoids. The remainder of this thesis will deal with the non-mevalonate pathway for the production of lycopene in *E. coli*.

Chapter 4

Systematic target identification

It was discussed earlier that a central goal of metabolic engineering is the improvement of cellular phenotype, such as metabolite overproduction, by the introduction of genetic controls. To this end, metabolic engineering efforts have considered the properties of the *overall metabolic network*, in sharp contrast to the single-gene focus that characterizes typical applications of genetic engineering. Due to the lack of extensive knowledge about molecular interactions and their kinetics, the dissection and optimization of metabolic pathways is an outstanding issue of central importance to metabolic engineering (Stephanopoulos, Alper, & Moxley, 2004). The focus in this chapter is on the improvement of lycopene production in *E. coli* through the identification of putative parameters and gene knockout targets through the use of stoichiometric models. While these stoichiometric models do not address the issues of kinetics or regulation, they can still be useful in elucidating putative parameters and identifying key gene targets and metabolic nodes of interest.

4.1 Stoichiometric Model

We address these issues here computationally and experimentally in the context of lycopene synthesis in *Escherichia coli.* Our computational search makes use of a stoichiometrically balanced, genome-wide bioreaction network of *E. coli* metabolism whose fluxes are computed such as to maximize cell growth yield in the framework of Flux Balance Analysis (FBA) (Edwards & Palsson, 2000; Segre, Vitkup, & Church, 2002). Although this model is genome-wide and global for most metabolic reactions, it is important to note that it is a strictly stoichiometric model, totally devoid of any kinetic or regulatory information. Consequently, targets identified by this model improve product synthesis solely on the basis of increased availability of metabolic precursors and cofactor balancing. This beneficial effect may be negatively impacted by non-predictive, adverse kinetic and/or regulatory effects.

We employed this formalism to investigate the effect of gene deletions, the most common means of introducing genetic perturbations, on lycopene production. The *E. coli* *i*JE660a GSM model (Reed et al., 2003) served as the basis for this stoichiometric network. Furthermore, the *crtEBI* operon was added to the model along with updated isoprenoid synthesis reaction details discovered after the formulation of this model (Adam et al., 2002; Hecht et al., 2001), as indicated in **Table 4.1**. Using this updated model, a total of 965 metabolic fluxes (included exchange fluxes) were calculated such as to: (a) balance the rates of synthesis and depletion of 546 metabolites, (b) maximize cell growth yield subject to a Minimization of Metabolic Adjustment (MOMA) alteration for suboptimal systems, and (c) utilize glucose as the sole carbon source (Edwards & Palsson, 2000; Segre, Vitkup, & Church, 2002). When multiple enzymes encode the same

reaction (as is the case with isoenzymes), all instances of that reaction were removed from the stoichiometric matrix. To avoid selecting mutants with extremely low growth, a minimum growth requirement of 5 – 10 % of the maximum was enforced. Knockout candidates were compared on the basis of predicted production level after invoking the growth requirement.

Pathway	Gene	Reaction
Isoprenyl-pyrophosphate synthesis pathway	*dxs*	T3P1 + PYR → DXP + CO_2
	ispC	DXP + NADPH ↔ MEP + NADP
	ispD	MEP + CTP → CDPME + PPI
	ispE	CDPME + ATP → CDPMEPP + ADP
	ispF	CDPMEPP → MECPP + CMP
	ispG	MECPP ↔ HMBPP
	ispH	HMBPP → 0.5 IPPP + 0.5 DMPP
Isoprenyl pyrophosphate isomerase	*idi*	IPPP ↔ DMPP
Farnesyl pyrophosphate synthetase	*ispA*	2IPPP → GPP + PPI
Geranyltranstransferase	*ispA*	GPP + IPPP → FPP + PPI
Octoprenyl pyrophosphate synthase (5 reactions)	*ispB*	5 IPPP + FPP → OPP + 5 PPI
Undecaprenyl pyrophosphate synthase (8 reactions)		8 IPPP + FPP → UDPP + 8 PPI
Lycopene Pathway	*crtE*	IPPP + FPP → GGPP + PPI
	crtB	2 GGPP → PHYTO + PPI
	crtI	PHYTO + 8 NADP → LYCO + 8 NADPH

Table 4.1: Modifications of the *i*JE660a model. The following reactions were added to the stoichiometric model to account for lycopene production. Metabolite abbreviations are included in **Table 4.2**.

Abbreviation	Metabolite
CDPME	4-diphophocytidyl-2-C-methyl-d-erythritol
CDPMEPP	4-diphosphocytidyl-2-C-methyl-2-phosphate-d-erythritol
DMPP	Dimethylallyl pyrophosphate
DXP	1-deoxy-d-xylulose-5-phosphate
FPP	trans, trans Farnesyl pyrophosphate
GGPP	Geranylgeranyl PP
GPP	trans Geranyl pyrophosphate
HMBPP	1-hydroxy-2-methyl-2(E)-butenyl-4-diphsophate
IPPP	Isopentyl pyrophosphate
LYCO	Lycopene
MECPP	2-methyl-d-erythritol-2,4-cyclodiphosphate
MEP	Polyol 2-C-methyl-d-erythritol-4-phosphate
OPP	trans Octaprenyl pyrophosphate
PHYTO	Phytoene
PPI	Pyrophosphate
PYR	Pyruvate
T3P1	Glyceraldehyde 3-phosphate
UDPP	Undecaprenyl pyrophosphate

Table 4.2: Metabolite abbreviations for the modified *i*JE660a model. The reactions listed in **Table 4.1** utilize the following metabolite abbreviations which were either created or used from the original model.

4.2 Identification of putative parameters

Utilizing the tools of FBA with the combination of the MOMA addition for suboptimal systems, characteristic phenotype behavior can be extracted for the carotenoid system. Before investigating the impact of gene knockouts, the formalism of FBA may be used to help elucidate key putative parameters influencing lycopene titers.

4.2.1 Growth rate and glucose uptake

Initial simulations using FBA resulted in determining the relationship between growth and glucose uptake rates and the molar yield of lycopene. **Figure 4.1** illustrates typical simulation data indicating that maximal lycopene yield is achieved at a decreased growth rate with a high glucose uptake rate. In this case, two parameters were imposed: glucose uptake rate was set as a fixed constraint for the model, whereas the desired growth rate was set through imposing a ceiling for the value of growth rate in the constraints. While solving this problem subject to the maximization of growth, it is possible to deduce the influence these variables have on lycopene yield. Furthermore, these results indicate that the maximum, theoretical biocatalyst yield of glucose to lycopene by *E. coli* is approximately 0.31 g lycopene / g glucose, in the absence of growth.

4.2.2 Formate production

Beyond growth rate analysis, the formalization of FBA may be used to determine the impact of byproduct formation on lycopene yield. This impact may be assessed by

constraining the output of a byproduct pathway to varying values and assessing the impact on lycopene production. Such an analysis was conducted for several byproducts including acetate and ethanol and most resulted in an inverse linear relationship illustrating that as carbon was diverted to the byproduct, less lycopene can be formed. However, the analysis of lycopene production as a function of formate production showed a two-phase behavior, especially at reduced growth rates. **Figure 4.2** illustrates that formate production is indeed inversely proportional to lycopene yield. However, at lower growth rates, formate production is reasonably tolerable up to a critical threshold between 5 and 10 mmol formate/hr. After this level, formate production is strictly competitive with lycopene production. These results highlight the fact that formate should be reduced, and raise the possibility of a unique metabolic function revolving around the formate node of metabolism.

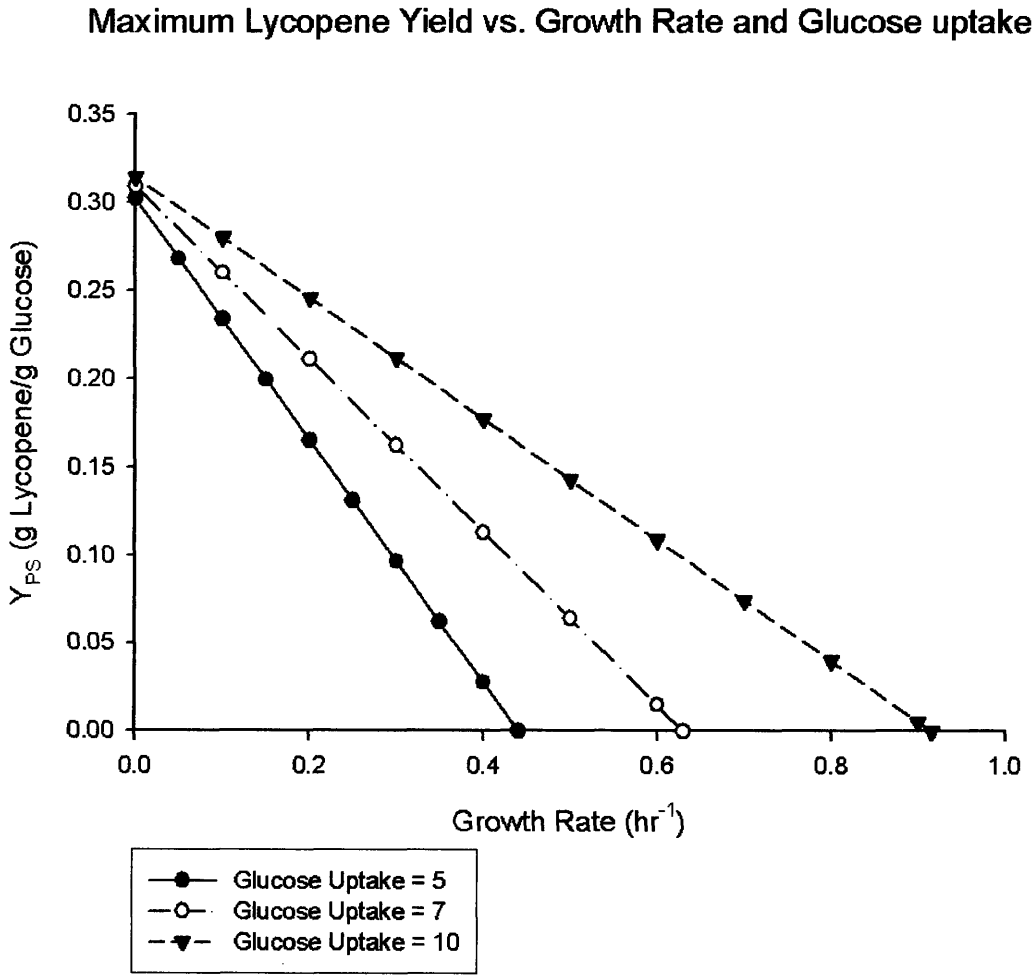

Figure 4.1: Impact of growth rate and glucose uptake rate on lycopene stoichiometric yield. A stoichiometric analysis of lycopene yield highlights that lycopene is produced at higher yields when cell growth is reduced and glucose uptake rates are maintained high.

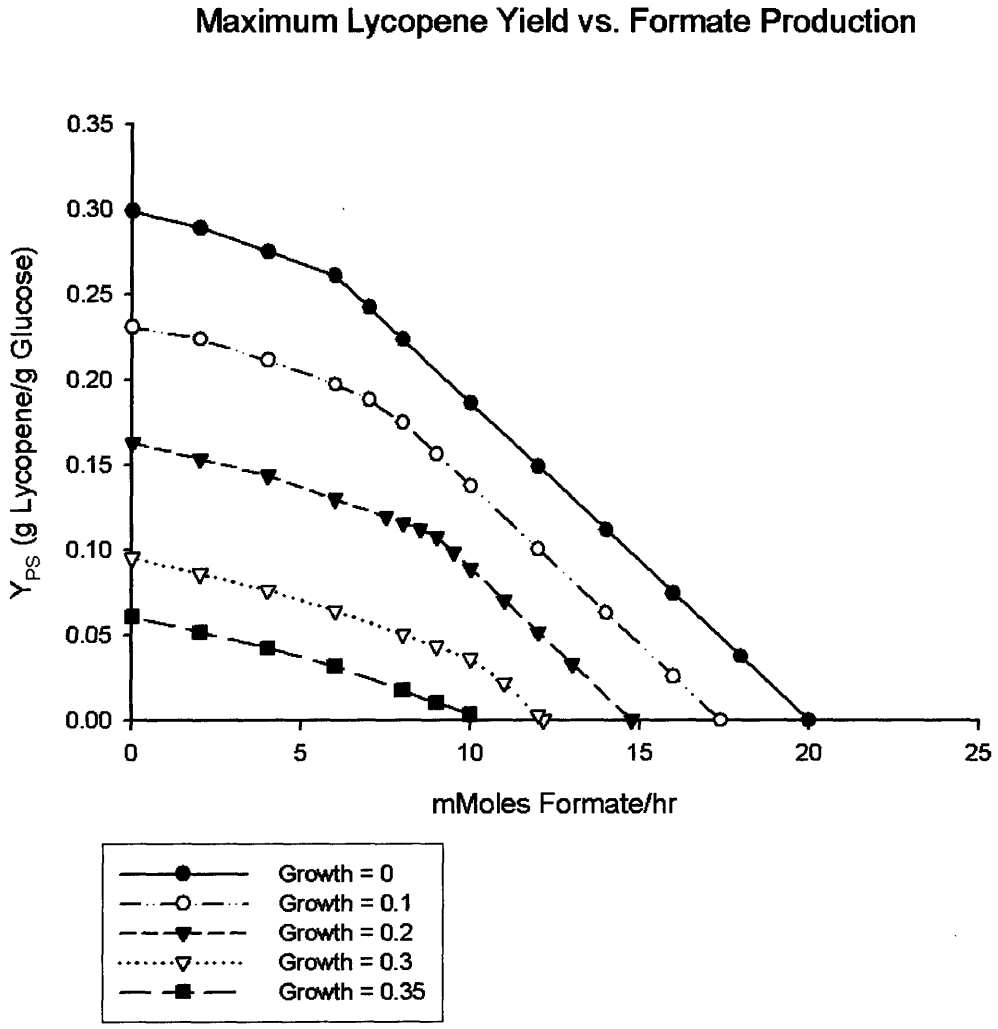

Figure 4.2: Impact of formate production on lycopene stoichiometric yield. Formate is a competitive byproduct to lycopene yield by diverting carbon away from the desired molecule. However, it is interesting to note two phases in this relationship between formate and lycopene yield. Formate production is relatively tolerable at production levels lower than 5-10 mmoles per hour. Above this level, formate is strictly competitive, especially at reduced growth rates.

4.2.3 Oxygen uptake rate

A stoichiometric analysis suggests that the lycopene yield in *E. coli* from glucose increases with oxygen uptake rate, presumably due to the large energetic requirement of lycopene production (8 CTPs and 8 ATPs per mole). **Figure 4.3** illustrates the relationship between oxygen level and the stoichiometric ycopene yield.

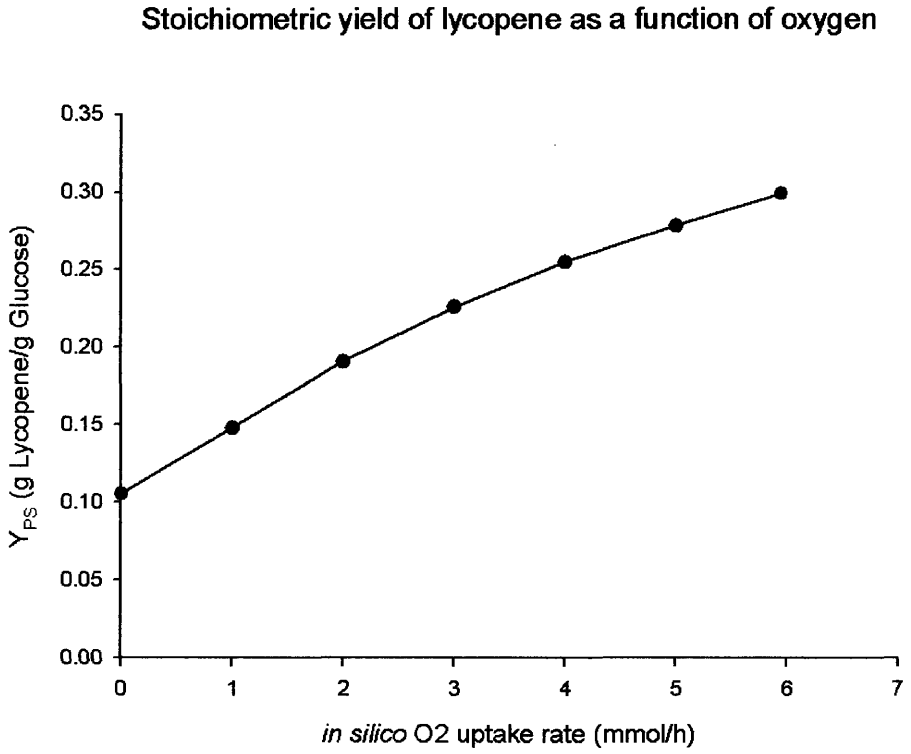

Figure 4.3: Impact of oxygen on lycopene stoichiometric yield. An *in silico* analysis using global stoichiometric models indicates that the maximum stoichiometric yield (g lycopene / g glucose) increases as a function of the oxygen uptake rate. In this calculation, the glucose uptake rate was set at 5 and the maximum yield of lycopene was calculated (thus at a growth rate of zero).

4.2.4 Carbon source optimization

Carbon source optimization is often an elaborate, trial-and-error based experiment for most bioprocess applications. However, the infrastructure of FBA allows for the quick evaluation of single and complex carbon sources. While glucose is often a carbon source of choice due to economic constraints, it may not be the best for a given bioprocess. A number of carbon sources were used as feed sources in the FBA simulation for lycopene production, which may be easily compared to experimental results. While glucose was a good carbon source, the simulation suggested that trehalose would give a higher yield, while glutamate was a worse carbon source. These comparisons are shown in **Figure 4.4** which illustrates how FBA may be used to preliminarily evaluate the potential for varied carbon sources. Furthermore, this figure is juxtaposed with experimental validation of these carbon sources which relatively support the findings of the stoichiometric analysis.

4.2.5 Summary

In general, these simulations revealed important general trends and relationships between lycopene yield and controllable factors. Therefore, these relationships suggest the need to reduce the growth yield, maintain a relatively high glucose uptake rate, maintain aerobic conditions, and minimize byproducts, especially formate to support enhanced lycopene production. Many of these parameters can be controlled through the

optimization of bioreactor design parameters and control strategies. Furthermore, these fundamental relationships serve as underlying principles behind selected gene knockout targets described in the next section.

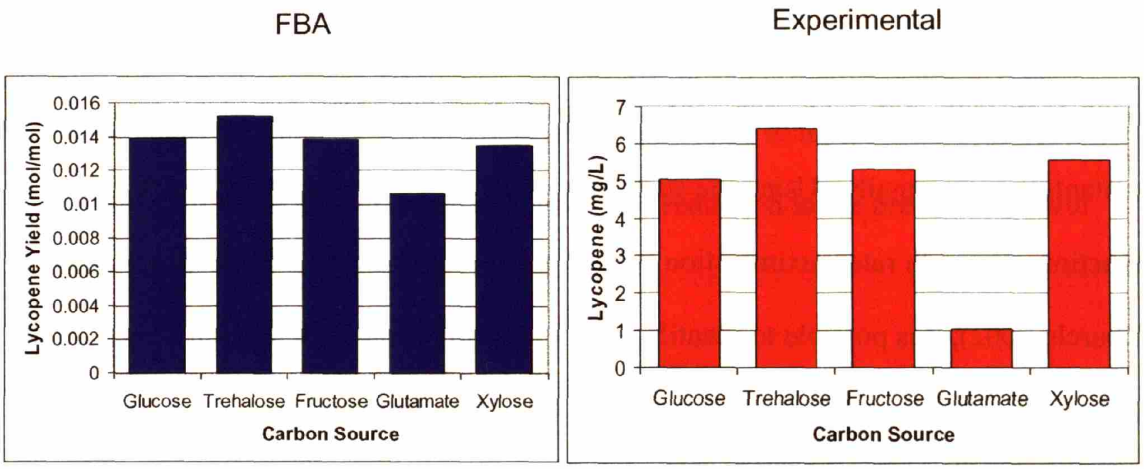

Figure 4.4: Carbon source optimization. Lycopene yield was analyzed as a function of carbon source. These results juxtapose *in silico* results (left, blue bars) with experimental results (right, red bars), which highlights a general consensus between experimental and computational approaches.

4.3 Single gene knockout targets

Beyond defining putative relationships between growth rates or byproducts and product formation, global stoichiometric models may be used to identify gene knockout targets (Burgard, Pharkya, & Maranas, 2003; Segre, Vitkup, & Church, 2002). As such, it is possible to simulate a knockout phenotype of every possible gene in the genome in very short computational times compared with the arduous task of creating each of these mutants experimentally. Using the standard stoichiometric constraints and objective function of growth rate maximization subject to the MOMA constraint (Segre, Vitkup, & Church, 2002), it is possible to identify gene knockouts which will naturally result in an increase in lycopene yield due to a rearrangement of the bioreaction network to favor different modes of cofactor regeneration or precursor balancing. Assessing the performance of all possible gnee knockout targets in the stoichiometric model yields a genome-scan, from which targets may be identified. It is possible to place a constraint of minimum growth to allow for the selection of experimentally feasible gene knockout targets.

4.3.1 Single knockout genome scan

Using the stoichiometric model along with a maximum growth objective function subject to a MOMA alteration, *in silico* genome-wide gene knockout simulations were conducted. The phenotype of specific gene knockouts was simulated by deleting the corresponding enzyme (i.e., reaction) from the stoichiometry matrix and calculating the resulting flux profile. When multiple enzymes encode the same reaction (as is the case

with isoenzymes), all instances of that reaction were removed from the stoichiometric matrix. To avoid selecting mutants with extremely low growth, a minimum growth requirement was enforced. Knockout candidates were compared on the basis of predicted production level after invoking the growth requirement. **Figure 4.5** summarizes the results of this genome scan for single gene knockout mutants. This genome scan identifies eight single gene knockouts which would produce a higher yield of lycopene by direct enhancement of the lycopene pathway and, indirectly, by lowering growth yield. **Figure 4.5** illustrates that most knockouts are not predicted to increase lycopene yield. Furthermore, all of the candidate knockouts show a reduction in the predicted growth yield.

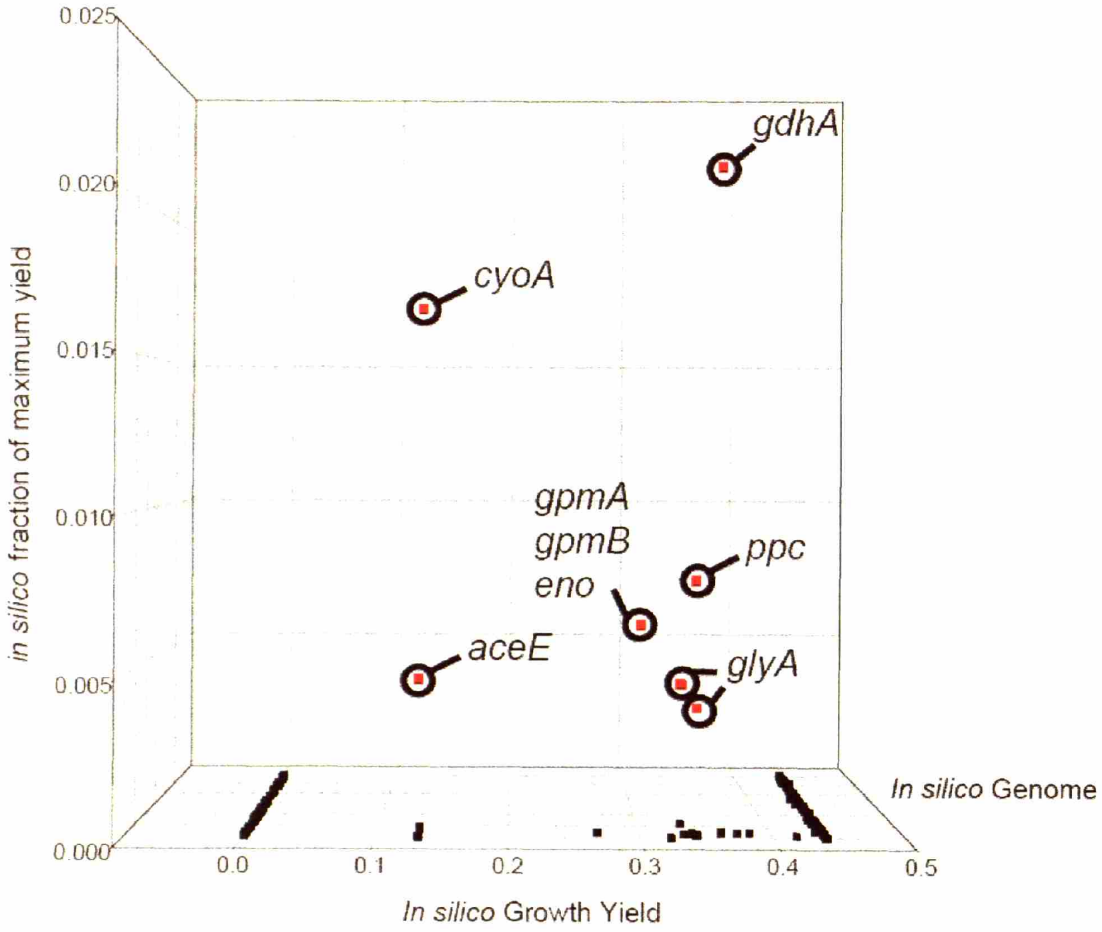

Figure 4.5: *E. coli* genome scan for single gene knockout targets. The phenotype of every possible single gene knockout was simulated using FBA with MOMA as an additional constraint. The above genotype-phenotype plot illustrates the effect of single gene deletions on lycopene yield as measured by the fraction of the stoichiometric maximum yield. These eight selected targets are discussed in further depth in the text.

As shown in **Figure 4.5**, a single knockout scan predicted eight genes whose deletion yielded enhanced product synthesis while satisfying a minimum growth requirement. The gene *glyA* appears twice since its function can be classified as both amino acid biosynthesis and vitamin/cofactor metabolism. The enzymes encoded by these genes are as follows: *aceE* (Pyruvate dehydrogenase), *cyoA* (Cytochrome oxidase bo3), *eno* (enolase), *gdhA* (Glutamate dehydrogenase), *glyA* (Glycine hydroxymethyltransferase), *gpmAB* (Phsophoglucomutase), and *ppc* (Phosphoenolpyruvate carboxylase). Of the eight predicted gene targets, two were eliminated from further consideration: *glyA*, due to a very low predicted improvement and *cyoA* that has been shown to exhibit a limited range of substrate utilization (Au, Lorence, & Gennis, 1985). After these exclusions, *gdhA*, *gpmA*, *gpmB*, *aceE* and *ppc* were selected as candidates for experimental validation. While *eno* also appeared as a candidate, it was not selected since the predicted phenotype was similar to the *gpm* isoenzymes and no prior strain containing the single knockout of *eno* was found in a literature search. Furthermore, while *gpmA* is the more prevalent isoenzyme form of the phosphoglycerate mutases in *E. coli* during the growth phase (Fraser, Kvaratskhelia, & White, 1999), the actual function and interaction of all phosphoglycerate related genes has not been fully determined. Additionally, *ppc* knockouts were found to be not viable in non-supplemented glucose-based media (McAlister, Evans, & Smith, 1981). Furthermore, *in silico* predictions indicated a reduced growth phenotype for each of these knockouts ranging between 40% and 75% of the maximum yield.

4.3.2 Linking of gdhA and NADPH

Of the five selected genes, all but *gdhA* apparently directly impact the supply of lycopene precursors while the *gdhA* knockout appears to increase the availability of NADPH, an important cofactor for lycopene synthesis required in a 16:1 mole ratio. An analysis of the predicted fluxes for NADPH consuming and generating reactions highlights a critical role for this cofactor. **Figure 4.6** presents a pie chart for the distributions of NADPH consuming reactions in a wild-type cell compared with **Figure 4.7** which presents the same chart for a *gdhA* deletion mutant. The amount of NADPH committed for glutamate metabolism is reduced from 47% of all NADPH in the wild-type to only 39% in the mutant. Additionally, a higher fraction of the NADPH in the mutant is committed for other pathways such as vitamins and lycopene. A higher transhydrogenase flux is predicted for the mutant. Furthermore, **Figure 4.8** illustrates that more NADPH is produced in a mutant cell, which collectively point to an increased availability of this important cofactor in a *gdhA* deletion mutant.

Overall, the results of this genome scan are consistent with existing strategies for increasing secondary metabolite production aiming at the reduction of byproduct formation, balancing of precursors and, in some cases, lowering the growth rate.

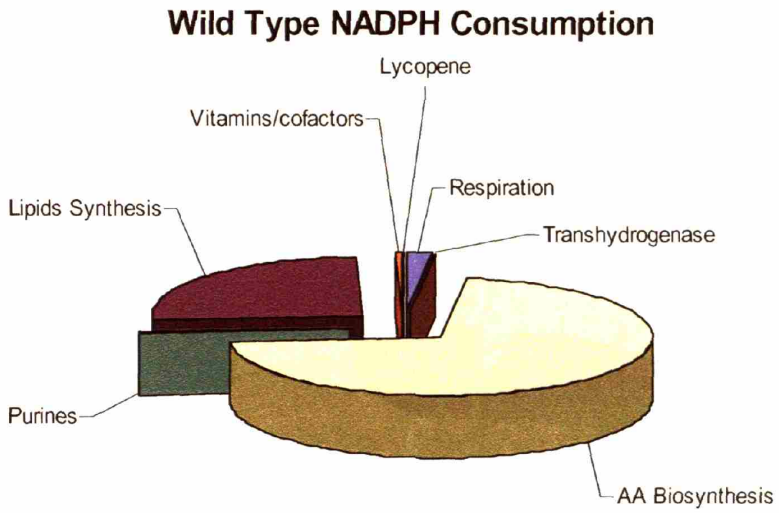

Figure 4.6: ***in silico*** **NADPH utilization in a wild-type strain.** In a wild-type strain, most of the NADPH pool goes to the formation of amino acids with 47% of NADPH consumption being attributed to the glutamate pathway.

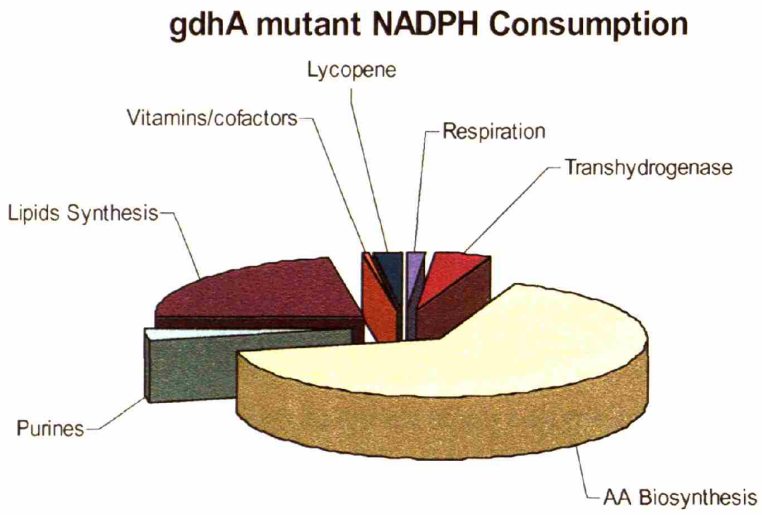

Figure 4.7: *in silico* NADPH utilization in a *gdhA* knockout strain. In a Δ*gdhA* strain, only 39% of NADPH consumption is being attributed to the glutamate pathway. Furthermore, a larger percentage of NADPH is predicted to be used for lycopene production and transhydrogenase activity compared with the control in **Figure 4.6**.

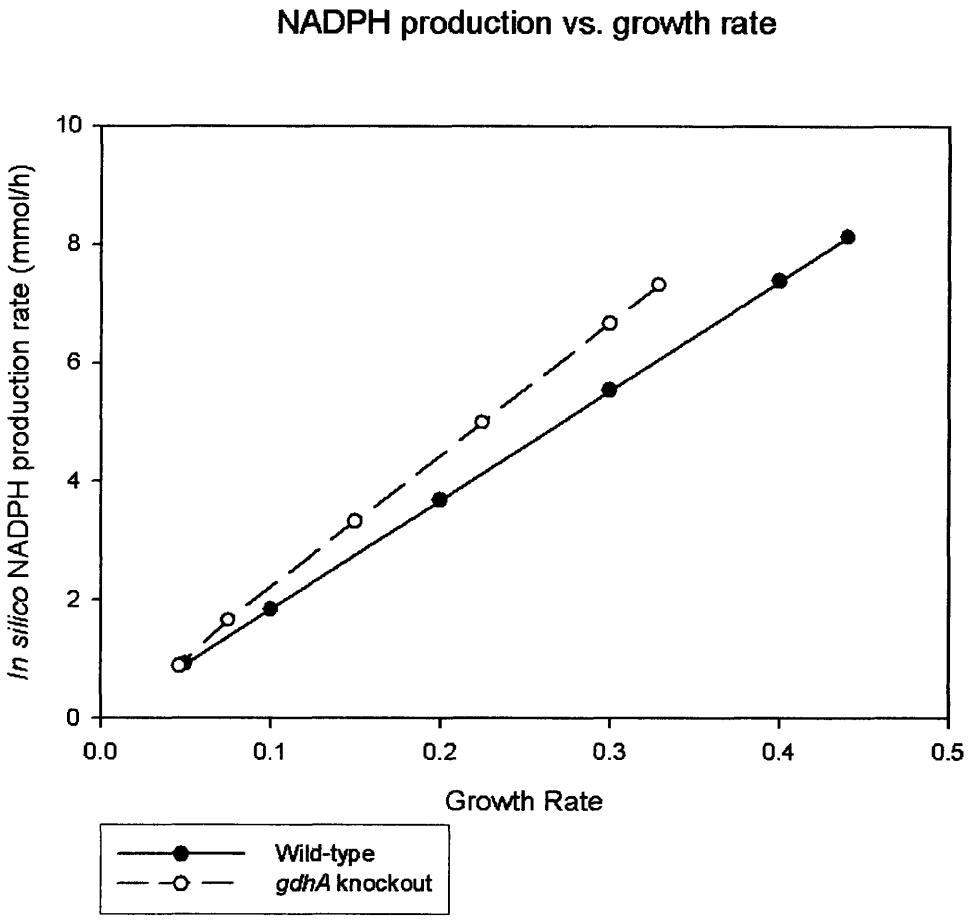

Figure 4.8: NADPH production rates. The *in silico* rate of NADPH production is predicted to be higher at nearly all growth rates for the *gdhA* mutant strain compared with the control. This increased rate together with a smaller percentage of NADPH used by amino acids and glutamate production account for the predicted increased NADPH availability in the mutant strain, and thus the predicted increase in lycopene yield.

4.4 Multiple gene knockout targets

Optimization of a secondary-metabolite phenotype, such as lycopene production, obviously depends on the modulation of several genes. Hence, multiple gene knockouts need to be similarly evaluated. The difficulty here is that exhaustive investigation of all possible gene knockout combinations leads very quickly to combinatorial explosion: ${}_{965}C_2$ combinations of all possible double mutants, and so on. Hence, *sequential* and iterative optimization approaches are often invoked whereby single gene knockouts are investigated in the genetic background of deletion mutants identified for their improved phenotype from previous iterations. Such procedures emulate optimization routines of the type of steepest descent for non-linear optimization problems. However, while properties of continuity and convexity assure a certain degree of success in the solution of mathematical problems, no such properties have been demonstrated for metabolic networks. Consequently, there can be no assurance about the results of such sequential optimization procedures.

4.4.1 Multiple knockout identification

We first investigated multiple gene knockout mutants following a sequential approach: a gene was first identified whose deletion yielded maximum lycopene improvement and double mutants were subsequently sought by scanning the effect of additional gene knockouts in the genetic background of the single gene knockout, and so on for higher mutants. Certain combinations of gene knockouts yielded extremely reduced growth phenotypes *in silico*, so a growth rate minimum was required of all

mutants equal to 5% of the maximum, wild-type prediction. **Figure 4.9** summarizes the results of several multiple knockout constructs of considerable interest. As a double knockout construct, *gdhA*/*gpmA* or *gdhA*/*gpmB* is predicted to outperform the other candidate combinations. However, all triple knockout constructs based upon *gdhA*/*gpmA* or *gdhA*/*gpmB* are predicted to have an extremely low growth rate (less than the 5% threshold), which warrants their removal from further consideration on the basis of the minimum growth rate requirement despite their predicted high product yield. Additionally, *talB* and *fdhF*, although absent as single knockout candidates, become key gene targets in the construction of double or triple knockout mutants. Of further interest, the gene *talB* is predicted to improve production in a *gdhA*/*aceE* background, yet it decreases the yield in the *gdhA* background when the enzymatic activity of *aceE* is present. **Figure 3.1** depicts the reaction network along with these candidate gene targets. It underlines the rationale for specific combinations, such as *aceE* followed by *fdhF* as a knockout scheme. In this case, the knockout of *aceE* would presumably increase formate production, whose flux may then be redirected through an *fdhF* knockout. Furthermore, the necessity to reduce formate is one of the putative factors identified using the stoichiometric model in Section 4.2.2. These features illustrate the need for the invoked systematic approach to identify gene knockout targets.

A final issue involves determining an endpoint for this analysis. The path of maximum phenotype increase was followed to predict a quadruple knockout mutant. However, this resulted in the selection of *ppc* as the next target, which is infeasible in a glucose-based medium. Furthermore, this mutation was predicted to impart only a marginal increase in the overall lycopene yield.

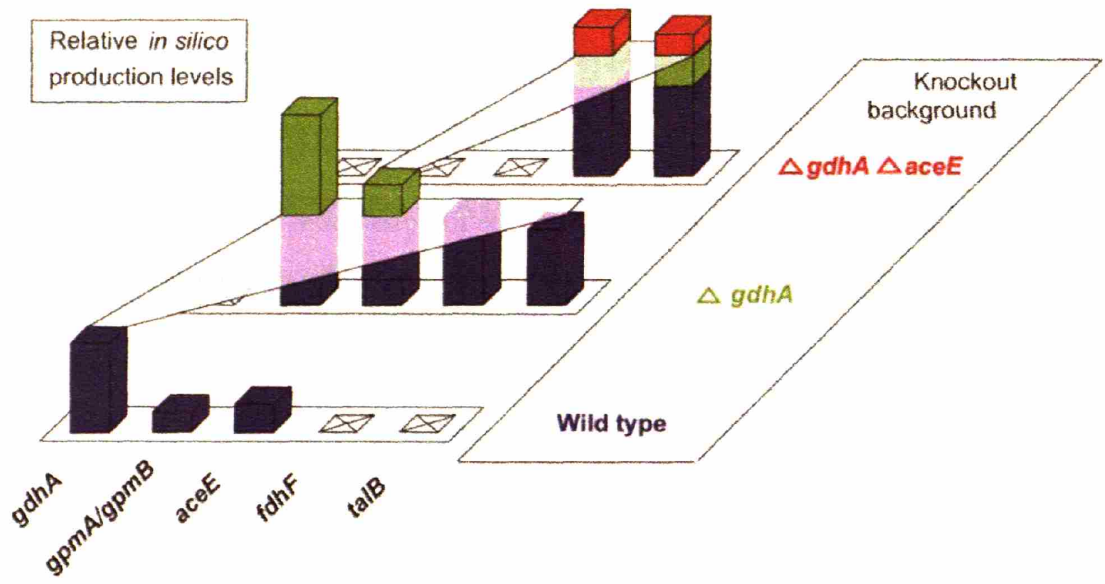

Figure 4.9: Identification of sequential, multiple gene knockout targets. The process of maximal sequential phenotype increase is illustrated. The production yields for each genetic background are simulated similarly to the method followed in Figure 2, but in a different genetic background for the starting strain. A triple knockout construct based on the double mutant *gdhA*/*gpmB* was excluded as it violated the minimum growth rate requirement. On the other hand, predicted triple constructs in *gdhA*/*aceE* background continue to show an increase in lycopene yield. The path of maximal phenotype increase is given by the solid lines. However, since a *gdhA*/*gpmB* knockout has been excluded for triple knockouts consideration due to growth rate, the next highest optimal path is followed. These results indicate that novel gene targets arise as the genotype is altered as result of gene knockouts. This is especially evident in the case of *talB*. Although *talB* increases the production level in a *gdhA*/*aceE* knockout background, it is detrimental in a *gdhA* only knockout background.

4.4.2 Sequential vs. simultaneous searches

Search strategy is an important consideration when approaching the problem of selecting multiple gene targets, as will be discussed in more depth in future sections and chapters. In the previous section, an iterative, sequential approach was used to select for subsequent gene knockout targets. However, following this path of steepest ascents to reach the global maxima depends on a number of factors about the shape of the optimization function. To address these issues, we sought to compare the above results to those obtained by an exhaustive investigation of all possible double mutants. An exhaustive simulation is computationally expensive, however provides insight into the topology and behavior of the metabolic space. At the end of this simulation, the highest yielding viable knockout (two *cyoA* combinations were excluded, as this gene was eliminated previously) was predicted to be a *gdhA*/*gpmA* or *gdhA*/*gpmB* construct which is similar with the result obtained using the sequential approach. Additionally, this analysis predicted that most double gene knockouts would not significantly increase the calculated lycopene yield.

Following the path of maximal sequential phenotype increases for predicting yields of double knockout constructs cannot, obviously, identify combinations of two synergistic genes which have no phenotype impact individually as single knockouts. As shown in **Figure 4.10**, of the top 90% phenotypes in the double knockout metabolic space, 98.6% contain at least one gene which elicits a high increase in lycopene yield as a single knockout. Additionally, all of the desirable phenotypes (those exhibiting the highest lycopene yield) reside in this subset of genes. Only 1.4% of the top 90% of

double knockout phenotype constructs would be unattainable following a sequential approach to target identification, however, the highest resulting phenotype in this subset of combinations is only 60% of the maximum yielding predicted double knockout construct. These results suggest that, *for this particular system,* the most desirable phenotypes are attainable using a sequential genome search strategy.

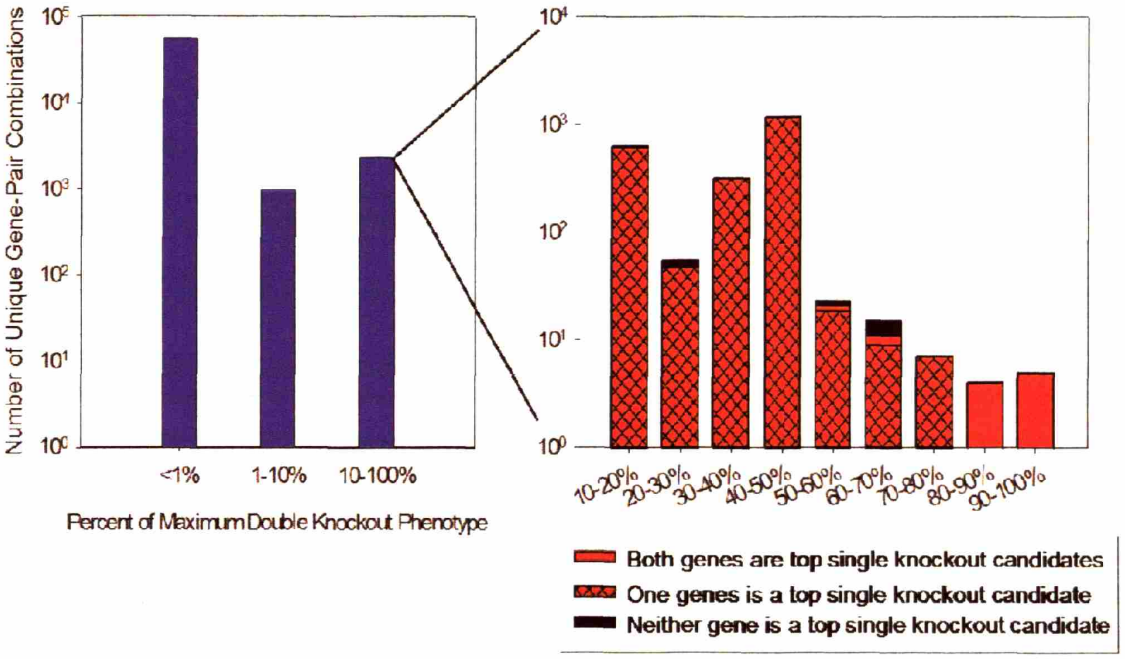

Figure 4.10: Simultaneous approach to multiple target identification. This figure depicts the results of the exhaustive double knockout search. In the first graph, it is evident that the majority of double knockouts have little impact on lycopene yield (A). The usual 5% minimum growth requirement has been imposed. Double knockout constructs are scored as a percent of the production level of the highest producing double knockout. The second graph (B) stratifies the top 90% of the double knockout phenotypes found by this exhaustive search. These results indicate that 98.6% of all possible combinations consist of either one or more genes from top candidates found from the single knockout illustrated in **Figure 4.5**. More importantly, all maximum phenotypes can be identified through the sequential search.

4.5 Experimental validation of targets

Gene knockout experiments were conducted along with shake-flask fermentations to experimentally test the predictions of the previous simulations. Knockout constructs were created using PCR product mediated inactivation (Datsenko & Wanner, 2000) in a recombinant *E. coli* strain already engineered to produce lycopene at high yields through chromosomal over-expressions of the *dxs*, *ispFD*, and *idi* genes. Since this strain simply contains chromosomal over-expressions of endogenous genes (which does not change the bioreaction network), the stoichiometric model used to identify gene targets is still suitable for this host. This strain was expressing the heterologous *crtEBI* operon on a pAC-LYC plasmid encoding for the additional genes required to produce lycopene (Cunningham FX Jr, 1994). **Table 4.3** summarizes the results from eleven so constructed knockout mutants and presents the lycopene production at the point of glucose exhaustion when strains were grown in an M9-Minimal media with glucose as the sole carbon source. In general, the experimental results validated the stoichiometric analysis and led to the formation of the best triple knockout construct comprising *gdhA*, *aceE*, and *fdhF* gene deletions. This strain increased lycopene yield by 37% after 15 hours compared with the pre-engineered control strain.

Knockout Construct	Growth Rate (hr^{-1})	Actual Percentage of Parental	Predicted Percentage of Parental	Percent Increase in Lycopene Content (PPM)
None	0.67	100%	100%	0% (4700 PPM)
Single Knockouts				
gdhA	0.55	82%	75%	13% (±4)
gpmA	0.44	66%	40%	-8% (±3)
gpmB	0.55	82%	40%	7% (±2)
aceE	0.52	78%	68%	9% (±4)
fdhF	0.57	85%	100%	4% (±3)
Double Knockouts				
gdhA, aceE	0.52	78%	56%	13% (±4)
gdhA, gpmA	0.37	55%	9%	12% (±3)
gdhA, gpmB	0.49	73%	9%	18% (±3)
gdhA, talB	0.46	68%	62%	3% (±4)
Triple Knockouts				
gdhA, aceE, talB	0.44	65%	44%	19% (±4)
gdhA, aceE, fdhF	0.38	56%	54%	37% (±3) (6600 PPM)

Table 4.3: Experimental results of single and multiple gene knockouts. Mutant growth rates and lycopene production (shown in ppm) are compared with the corresponding levels obtained in the non-mutated parental strain with zero knockouts. Growth rate data are compared as a percentage of the parental strain and juxtaposed with the predicted values. It is important to note the differing effects of the two gpm isoenzymes (gpmA and gpmB), as the knockout of gpmA appears to give the greater impact. Total lycopene content increases with multiple knockouts obtained along the path of highest production, with the exception of gpmA. Numbers in parenthesis indicate the standard deviations among replicate culture experiments. Different batches of medium caused the absolute value of lycopene production to vary slightly. As a result, all trials were conducted along with the parental strain as an internal control.

4.5.1 Conclusions from results

Five major conclusions arise by comparing this experimental data to the results of the simulations. First, the trend of actual mutant growth rates compares qualitatively to the predicted values. Second, there is continuing improvement of lycopene yield with an increased number of selective gene knockouts. This trend reflects the selection criteria applied in the identification of gene targets. Third, the *gdhA*/*gpmB* double knockout construct produced the highest yield among double knockouts at 18% above the parental strain, as predicted by the simulations. Fourth, following the path of highest product yield in combination with the minimum growth requirement yielded a triple knockout construct of *gdhA*/*aceE*/*fdhF* which produced the highest yield of 37% above that of the parental strain. Finally, gene targets selected as being important in triple knockout constructs, matched computational predictions to be either ineffective as a single knockout (*fdhF*) or detrimental as double a knockout (*talB* in *gdhA* background) where the advantage in lycopene production created by *gdhA* was reversed. Overall, the experimental results followed the trends suggested by the simulations.

One notable exception to the qualitative adherence of experimental data to computational results is the impact of the *gpmA* knockout. As a single knockout, this construct resulted in a decrease in lycopene yield, and was consistently lower than any *gpmB* construct. Since *gpmA* is a more dominant isoenzyme than *gpmB,* the expected metabolic consequences are higher and its impact is most evident in the substantial growth rate decrease in *gpmA* knockouts. Phosphoglycerate mutase knockouts could lead to the accumulation of 3-phosphoglycerate, which is known to have regulatory functions

within the cell, especially as it serves as a precursor to amino acids, and may be negatively interacting with lycopene production in this experiment. The results from this gene knockout construct illustrate that gene knockouts can increase precursor availability and lead to increased lycopene production *to the extent that regulatory effects elicited by the deletion of the gene* do not interfere with product synthesis.

Since the predictions were based on glucose as the sole carbon source, nonviable knockouts such as *ppc* were excluded from further analysis. Despite this fact, *ppc* was still found to impart an increase in lycopene yield compared to the parental strain when grown in minimal media with 0.3% Casamino acid supplementation. The impact of a *ppc* knockout was approximately a 20-25% increase over the value of the control.

4.5.2 Comparison to random perturbations

The significance of these results should be examined relative to the lycopene yield improvements afforded by the deletion of other genes not identified by the flux balance simulations. To this end, *libraries* of random genome transposon knockouts using the pJA1 vector (Badarinarayana et al., 2001) were constructed. Such randomized libraries do not show any significant increase in the overall lycopene yield, which is illustrated by **Figure 4.11**. The results of **Figure 4.11** should not be construed to imply that there is no random gene knockout that can increase significantly lycopene yield. Only a small fraction of all transposon strains were analyzed individually and, in fact, an efficient and more exhaustive screening or selection process could identify high yielding knockouts targeting critical regulatory elements within the cell. For the purpose of this comparison,

we examined 8 heterogeneous cultures and 25 randomly-selected, individual colonies of the random transposon mutagenesis library cultured separately in shaker flasks. None of these mutants compared favorably with the selected knockout mutants. In essence, these results indicate that the particular knockout targets identified bring about a measurable effect on lycopene accumulation that is above any lycopene change impacted by gene knockouts at random. **Figure 4.11** juxtaposes the results of the systematically selected gene targets with the random library strains showing that gene knockouts of targets identified through stoichiometric modeling perform better than average random gene knockouts. We note that it is not yet possible to compare multiple knockout constructs due to the inadequacy of currently available genetic tools to create multigenic knockout libraries. A properly guided sequential search provides an efficient approach to multigenic modifications required to produce a desired phenotype. It is important to note that these results are not meant to imply that any random gene knockout is inferior to the systematically designed strains. The identification of superior knockouts using this method of genetic perturbation is discussed in depth in the upcoming chapters.

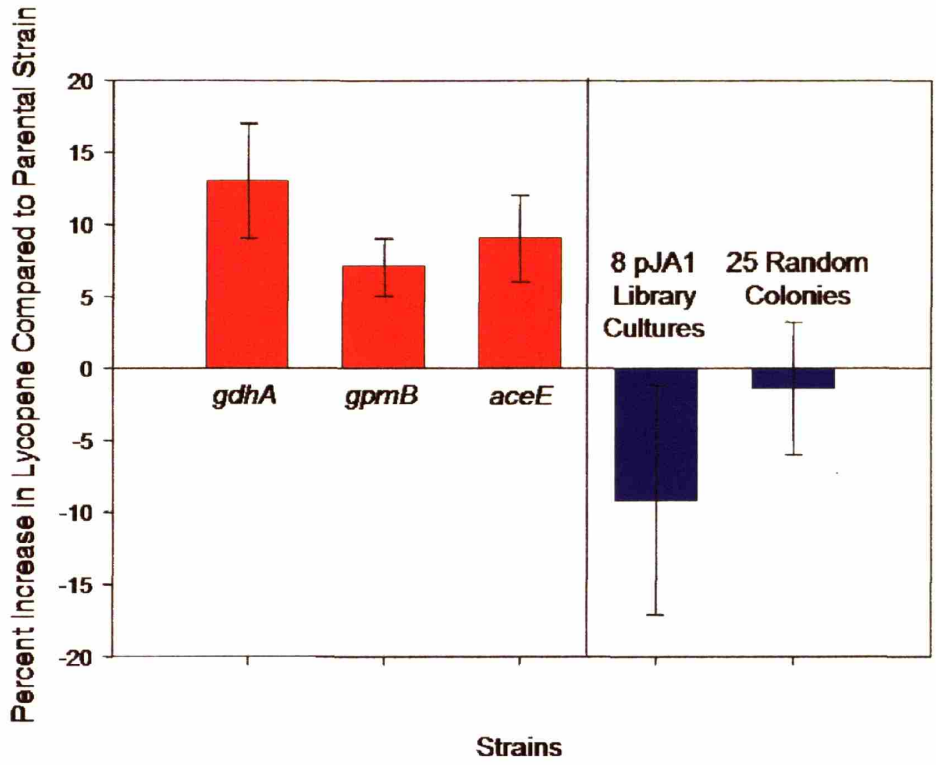

Figure 4.11: Comparison of selected mutants to random libraries of knockouts.

The results of the selected single knockout targets are juxtaposed to random transposons. In particular, 8 heterogeneous cultures and 25 individual colonies from the random transposon mutagenesis library were cultured separately in shaker flasks. Both the heterogeneous culture of a random transposon mutants and the average of the randomly selected colonies appeared to decrease the lycopene production. None of these cultures compared favorably with the systematically selected single gene knockout mutants of *gdhA*, *gpmB*, or *aceE*.

4.6 Summary

A systematic, exhaustive computational or experimental search of all feasible gene knockouts in *E. coli* to determine the genotype yielding the optimal lycopene production phenotype is a tedious and often difficult process. The effects of individual gene knockouts are not necessarily additive in determining the effect(s) of combinations of knockouts which successfully enhance a phenotype. As the search space increases to include all possible double or higher (triple) knockouts, systematic and exhaustive searches, both computationally and experimentally, become almost infeasible. This research suggests that this nonlinear process may be initially optimized, to ensure the proper supply of precursors, in a manner analogous to the method of steepest descent for nonlinear function optimization. Since this approach may often yield a local, as opposed to global production maximum, the results of other possible trajectories must be compared to identify the one with the most promising end point along the phenotype contour. In the method followed here, single gene knockout targets are first identified, of which the highest producer is selected. With this mutant as the new background, new knockouts are determined and the highest mutant is selected again. Through this process, knockout mutants are constructed with progressively increasing production phenotype.

This search technique generated novel single and multiple gene targets for increasing the production level of lycopene in *E. coli*. Furthermore, combinations of gene knockout targets for multiple knockout constructs were identified. The genes *talB* and *fdhF* exemplify the unique aspects extracted from stoichiometric modeling. As the cellular genotype changes, new stoichiometric targets arise. These new targets both

computationally and experimentally illustrate the intrinsic link between cellular genotype and phenotype. Single gene modifications may not be additive in nature and thus a more systematic analysis is required to extract the optimal combinations of gene modifications. Likewise, inferences about the impact of perturbations in one strain may not be immediately transferable to another strain possessing a modified genetic background. This difficulty is enhanced by the inability of current models to capture the regulatory effects which could negatively impact product formation.

An exhaustive genome scan for all possible double knockouts failed to provide any unique or interesting targets yielding desirable phenotype characteristics. This simulation also generated insights about the potential and limitations in transferring cellular information between strains of different genotypes. An undirected combination of the top single knockout gene targets could result in suboptimal combinations, yielding as low as 50% of the lycopene production in the predicted highest attainable phenotype. Additionally, the best phenotypes could be extracted by following a sequential phenotype enhancement, which saves computational and experimental efforts. Although these results cannot be generalized for other products and strains, they nevertheless underline the ability of sequential approaches to reach very interesting phenotypes in certain cases.

Most of the gene targets identified in this study could be superimposed on a simple network diagram modeling central carbon metabolism. Discussion of the need for global as opposed to local metabolic models has indeed been brought to the forefront of research efforts after genome sequencing and a refocusing on systems biology approaches to cellular systems. As such, it is difficult to *a priori* determine the size of a model needed to capture the behavior of a system. In fact, the key utility of these large,

global models is the extensive linking of distant reactions and metabolites through cofactor, energetics, and precursor balancing.

It should be further noted that this type of analysis is not limited to gene deletions only. It is possible to similarly explore other genetic modifications, such as gene expression amplifications, to identify putative parameters impacting cellular phenotype. As long as positive interactions exist of fundamentally stoichiometric nature, they can be uncovered by this approach generating additional promising genetic targets that can influence positively the cellular phenotype.

Neither flux balance analysis nor the *i*JE660a GSM accounts for genetic regulation and other possible cellular interactions. It is conceivable and quite possible that genetic regulations outweigh stoichiometric effects. As the former are notably absent from the stoichiometric model used, neither the latter nor alternative search methods can capture other possible gene targets arising from these complex regulatory interactions. When such advanced models of cell function become available, identifying optimal gene targets will still be a demanding undertaking that can be facilitated by the findings of this study. These more comprehensive models will allow a more detailed mapping of the phenotypic landscape along with a thorough evaluation of various search methods for promising genetic targets for metabolic engineering. Absent such models, a *sequential*, iterative optimization provides a reasonable and feasible alternative that can yield promising targets, as in the case of lycopene synthesis examined in this work.

Chapter 5

Mapping the metabolic landscape

Lycopene production in the systematically identified knockout strains described in Chapter 4 was still below the stoichiometric maximum allowed level, presumably limited by unknown kinetic or regulatory factors that are unaccounted in stoichiometric models. To investigate these factors and identify novel gene knockout targets, we undertook a global transposon knockout search to identify gene targets (hence forth referred to as *combinatorial* gene targets) that could further increase lycopene production in *E. coli*. By further combining these targets with the previously identified *stoichiometric* targets (from the stoichiometric modeling), we constructed and analyzed a total of 64 strains with different knockout genotypes which span the lycopene metabolic landscape. Combining these two distinct sets of gene knockout targets allowed, for the first time, the definition and visualization of the metabolic landscape, supporting valuable observations regarding strain improvement strategies. An analysis of the interaction of gene targets in this landscape provides invaluable insight into defining search strategies.

5.1 Identifying combinatorial targets

To identify additional knockout targets impacting the lycopene phenotype via regulatory, kinetic, or other unknown mechanisms, we undertook a global transposon library search in the background of the pre-engineered parental strain. A 10^5 library of random gene knockouts created through the use of the pJA1 transposon vector (Badarinarayana et al., 2001), was created in the pre-engineered parental strain overexpressing *dxs*, *idi*, and *ispDF*. Screening this transposon library on minimal media plates and revalidation of the phenotype ultimately identified three gene targets that correlated with lycopene over-production. Upon sequencing, these combinatorial targets were identified as *hnr* (also known as *RssB)*, *yjfP* and *yjiD*.

The gene *hnr* is a response regulator (part of the two component system) and is responsible for recruiting the proteolysis of the stationary phase sigma factor, σ^S (encoded by *rpoS*) (Muffler et al., 1996; Sandmann, Woods, & Tuveson, 1990). It is noted that *rpoS* has been implicated in the over-production of carotenoids, both in a heterologous setting in *E. coli* and in endogenous production in *Erwinia herbicola* (Becker-Hapak et al., 1997). The gene *yjfP* is a 249 amino acid protein which is currently not annotated, but has been putatively categorized as either a non-peptidase homologue (Rawlings, Tolle, & Barrett, 2004) or as a putative hydrolase (1^{st} module) (Serres et al., 2001) . Finally, *yjiD* is a 130 amino acid protein with an unknown function (Serres et al., 2001). For this target, the transposon was found to be inserted between the identified promoter region and the gene for *yjiD* and will henceforth be referred to as Δ_p*yjiD*. For *hnr* and *yjfP*, several strains were identified in which the transposon was located in various regions of the gene. However, in all *yjiD* mutants, the transposon site

was only found between the promoter region and the gene. **Figure 5.1** presents the lycopene production of the identified combinatorial gene knockouts with respect to the control at 15 hours of culturing. We note that none of the previously identified stoichiometric genes surfaced in the combinatorial transposon search due to the relatively high threshold of lycopene accumulation level imposed in the selection of candidate strains, as no single stoichiometric gene knockout provided increases above 15%.

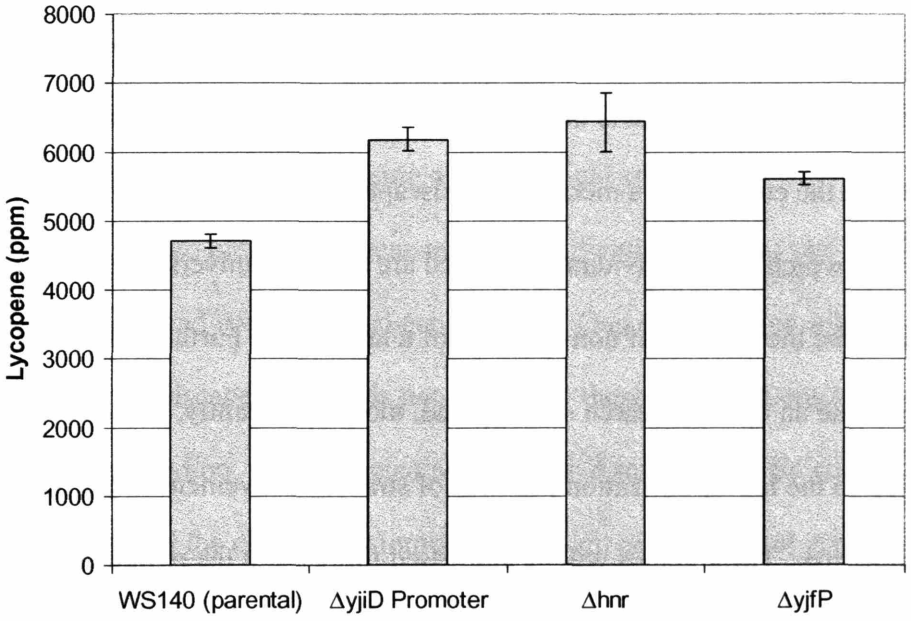

Figure 5.1: Comparison of combinatorial targets to parental strain. The three identified combinatorial gene knockout targets outperformed the parental strain at the 15 hour mark at an average of around a 30% improvement.

5.2 Mapping the metabolic landscape

There are many uncertainties associated with metabolic landscapes and search strategies. Using the information gained from the systematic and combinatorial gene knockout searches, it is possible to make a metabolic landscape by creating all possible combinations of the previously identified targets. There are several unknown questions about the topology and interaction of this landscape. In particular, these questions revolve around two main uncertainties about metabolic landscapes: (1) how non-linear are metabolic landscapes and (2) what is the optimal search strategy for traversing these landscapes. Through the creation of a metabolic landscape, it is possible to assess whether gene targets which were individually selected are actually universal targets and it is possible to determine the sources of non-linearity of a landscape. Furthermore, it is unknown how to create an optimal search strategy and, more importantly, whether there is a unique solution to the bio-optimization problem of strain improvement rather than a number of local maxima. These issues will be addressed throughout the remainder of this chapter and continued through further landscape exploration presented in Chapter 6.

5.2.1 Creating a systematic-combinatorial metabolic landscape

Stoichiometric gene knockout targets (discussed in Chapter 4) were identified using a global, stoichiometric analysis of *E. coli* metabolism. A total of 7 single and multiple gene deletions, (Δ*gdhA*, Δ*aceE*, Δ*gpmB*, Δ*fdhF*, Δ*gdhA* Δ*aceE*, Δ*gdhA* Δ*gpmB*, Δ*gdhA* Δ*aceE* Δ*fdhF*), were predicted and experimentally validated to increase lycopene production, presumably by increasing the supply of precursors and cofactors that are

materially important in the lycopene pathway. The left panel of **Figure 5.2** depicts the methodology followed for the determination of the gene targets in the rational construction of multiple knockout strains. These seven mutations along with the parental strain comprise the set of eight systematically designed genotypes.

The combinatorial approach identified 3 gene knockout targets depicted on the right panel of **Figure 5.2**. Using these three identified targets, it is possible to create a total of seven gene combinations of single, double, and triple combinatorial target mutations (Δ*hnr*, Δ*yjfP*, Δ_P*yjiD*, Δ*hnr* Δ*yjfP*, Δ*hnr* Δ_P*yjiD*, Δ*yjfP* Δ_P*yjiD*, and Δ*hnr* Δ*yjfP* Δ_P*yjiD*). These seven combinations along with the parental strain constitute the combinatorial strain set comprising a total of eight strains.

These two methods of gene target identification point to two disjoint sets of stoichiometric and combinatorial gene targets. One of the major unknown factors of such a metabolic landscape is that it is not clear how these targets interact when combined. To answer this question, we conducted an exhaustive study of the 64 strains comprising all combinations of the eight stoichiometric and eight combinatorial genotypes. These target genes were modified in the background of a recombinant *E. coli* strain already engineered to produce lycopene at high yields through chromosomal over-expressions of the *dxs*, *ispFD*, and *idi* genes. Each of the 64 strains was evaluated on the basis of lycopene production over the course of a 48-hour shake-flask fermentation process. The resulting production profiles provided the information needed for the complete mapping of the lycopene metabolic landscape.

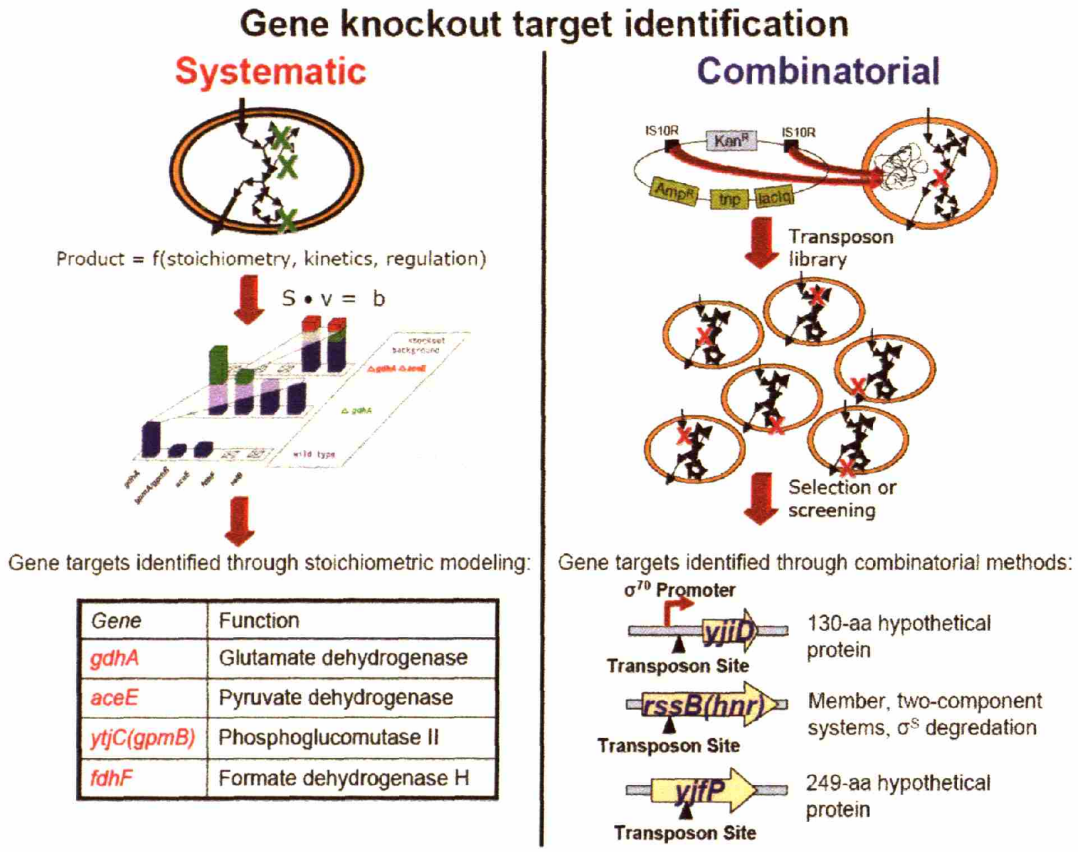

Figure 5.2: Systematic and combinatorial gene knockout target identification. Systematic targets (illustrated on the left) were identified through the use of global, stoichiometric modeling to identify gene knockouts which were *in silico* predicted to increase lycopene by increasing either cofactor or precursor supply. Combinatorial targets (shown on the right) were identified through the use of transposon mutagenesis. These targets were combined to create the unique set of 64 mutant strains used in this study.

5.2.2 Visualizing the metabolic landscape

The 64 strains comprising all combinations of the eight stoichiometric and eight combinatorial genotypes were analyzed through the use of 48 hour shake-flask fermentations. **Figures 5.3, 5.4 and 5.5** depict the landscape at 15 hours, 24 hours and the maximum lycopene over the entire 48 hours for each of the 64 mutant strains respectively. Several interesting observations arise from the topology of this surface, especially prevalent when viewing the maximum lycopene produced over the course of the fermentation, illustrated in **Figure 5.5**. First, two global maxima exist each with production levels around 11,000 PPM. The first strain contains the Δ*gdhA* Δ*aceE* Δ*fdhF* genotype, which is a purely stoichiometrically designed strain. The other maximum is Δ*gdhA* Δ*aceE* Δ_{P}*yjiD* which is created through the combination of stoichiometric and combinatorial targets. Second, several local maximum points are present with production levels ranging from 8,400 – 9,400 PPM, each formed from the combination of systematic and combinatorial targets. Third, the left quadrant of the graph indicates that the combination or stacking of more than one combinatorial knockout targets greatly *reduces* lycopene levels to below 2,000 PPM, and as low as only 500 PPM for some constructs. It is noted that 500 PPM is below even the production level of the wild-type strain of *E. coli* K12 devoid of up-regulations in the isoprenoid pathway.

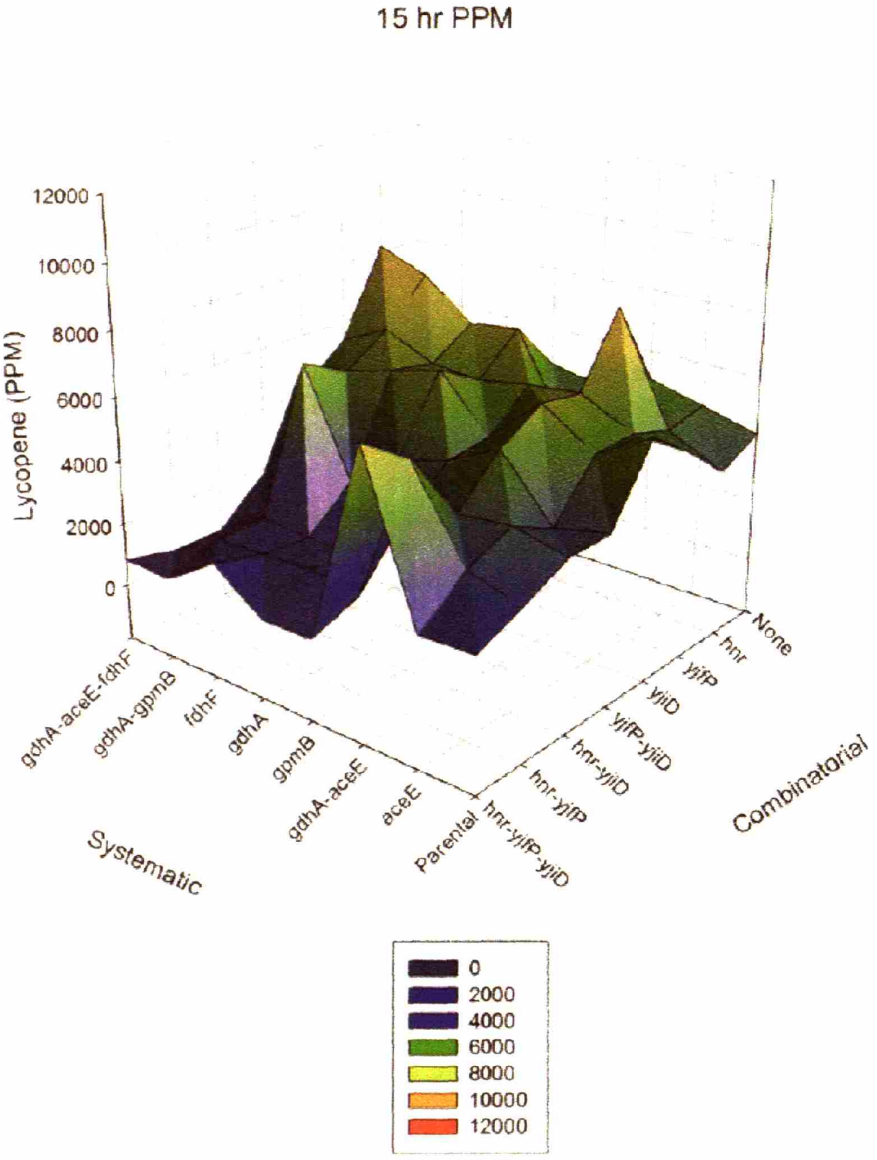

Figure 5.3: Visualization of the metabolic landscape at 15 hours. Lycopene values of the 64 mutant strains at the 15 hour timepoint are depicted in this metabolic landscape.

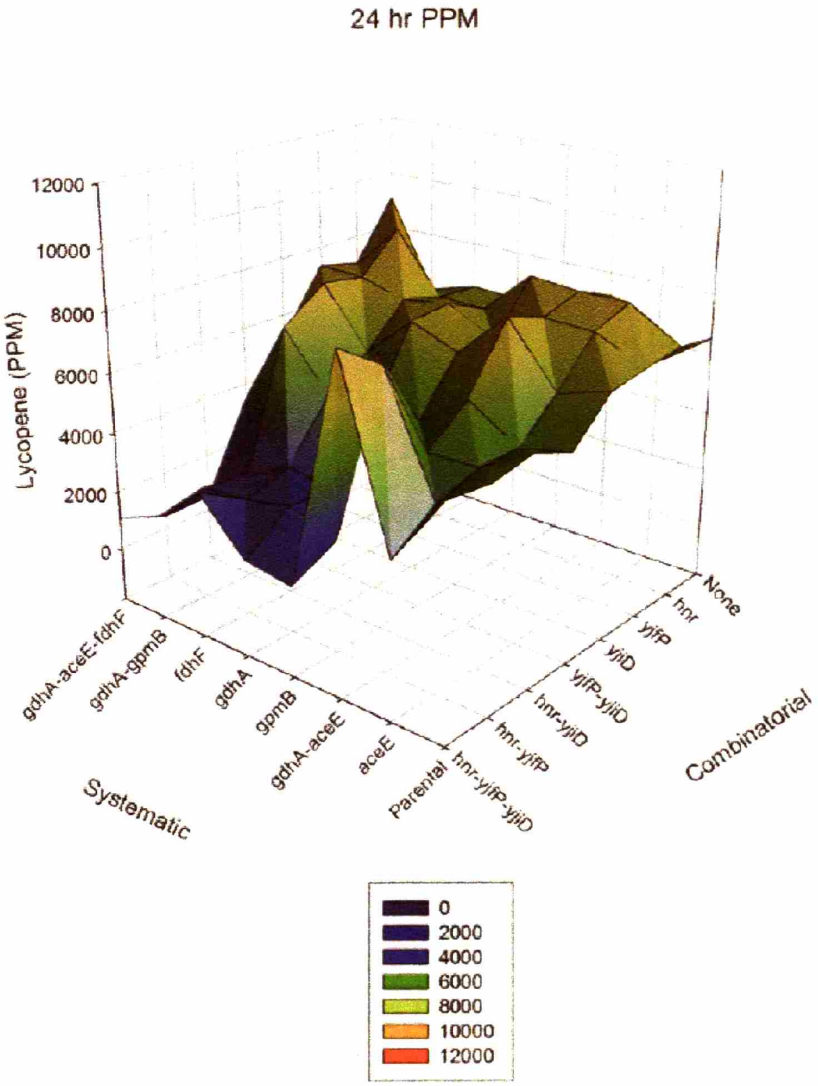

Figure 5.4: Visualization of the metabolic landscape at 24 hours. Lycopene values of the 64 mutant strains at the 24 hour timepoint are depicted in this metabolic landscape.

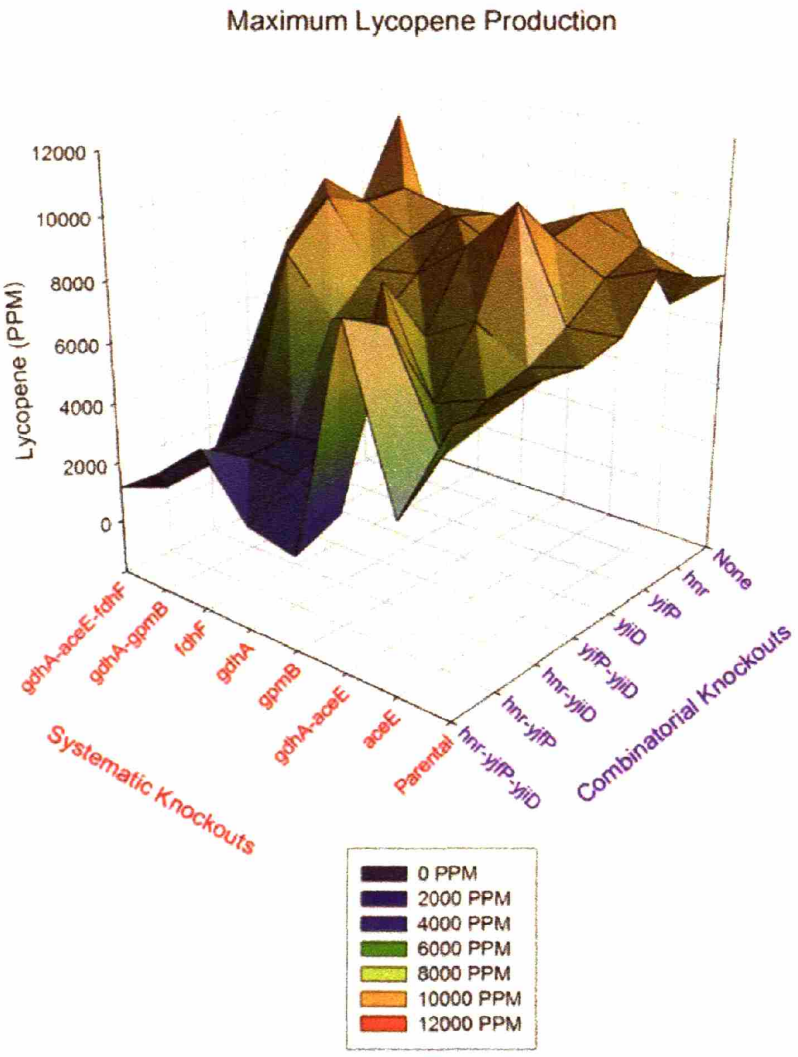

Figure 5.5: Visualization of the metabolic landscape—maximum production. The maximum lycopene production (in ppm) during the course of a 48 hour shake-flask fermentation is plotted. Among the interesting features of this landscape is the presence of two global maxima (at around 11,000 PPM) and several local maxima. Furthermore, certain combinations of combinatorial targets in most systematically-derived genetic backgrounds result in a substantial decrease in lycopene production.

5.3 Uncovering genetic interactions

Visual inspection of the landscapes presented in **Figures 5.3-5.5** suggests a highly nonlinear function with many local optima. Various statistical metrics may be used to parse out the impact of varying subsets of gene knockout targets to help uncover the underlying genetic interactions leading to the complexity seen in the landscape.

5.3.1 Impact of combinatorial targets

Initially, the impact of the seven combinatorial genotypes on the systematically designed strains was investigated. **Figure 5.6** presents a box-and-whisker type plot for assessing this impact. The boxes represent the average fold improvement for introducing the indicated combinatorial genotype into the background of each of the eight systematic genotypes. The whiskers highlight the minimum and maximum fold change elicited by the given combinatorial genotype over the eight genetic backgrounds of the systematic knockout strains. This impact was assessed at four different timepoints. The three, individual combinatorial targets, at least for the 15 hour timepoints, provided a consistant increase of lycopene production when placed in nearly any of the eight systematically designed strains (seen by boxes wich are above 1 with small whiskers, which are nearly all above the level of 1, with the exception of *hnr*). On average, each of the three combinatorial gene knockouts provide an increase to each of the eight systematic backgrounds as seen by boxes (representing average change) above 1 and small whiskers. In contrast, the stacking of combinatorial genes to make double and triple constructs is extremely detrimental as all boxes are below 1 for these constructs, indicating that the

addition of these double and triple knockouts to the background of any of most of the eight systematically designed genotypes reduces lycopene production. However, the impact of these stacked targets is still quite specific to the genotype as seen by the large whiskers, illustrating that the addition of certain stacked, combinatorial targets in a specific systematically-designed genotype can further increase the lycopene produciton. These results highlight the introduction of a non-linearity into the metabolic landscape from the combinatorial targets which are often of unknown or regulatory function.

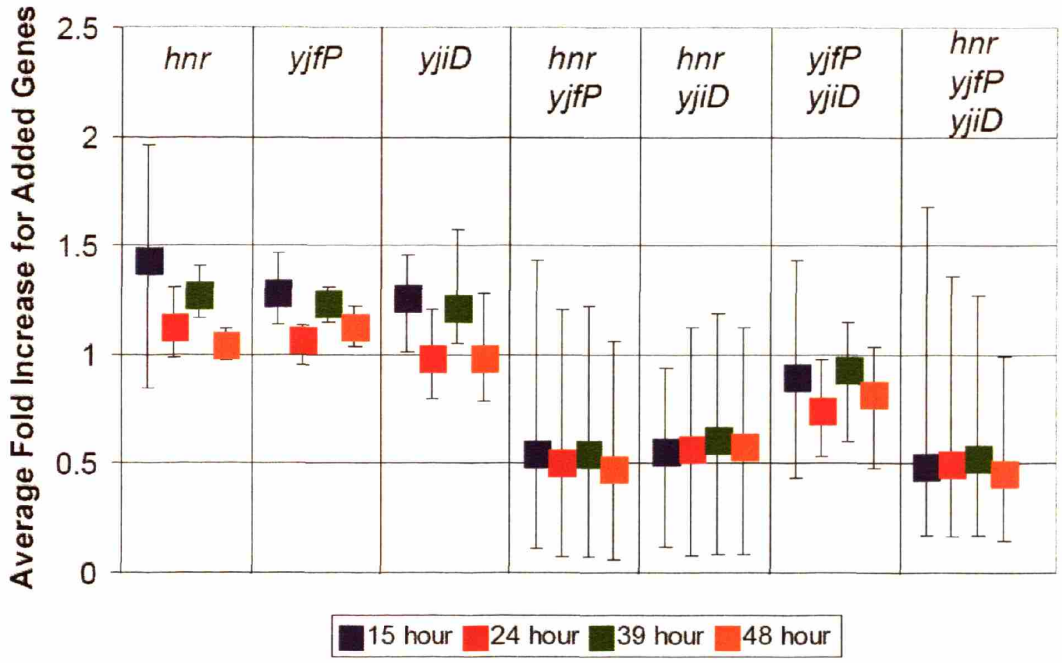

Figure 5.6: Impact of combinatorial genotypes on systematic backgrounds. This box and whisker-type plot illustrates the impact of the various combinatorial backgrounds in each of the eight systematic knockout backgrounds at 4 timepoints. In this case, boxes present the average fold increase in eight different strains created by the given combinatorial gene knockout placed into each of the eight systematic strains. While the boxes illustrate the average impact, the whiskers show the maximum and minimum impact of the combinatorial gene knockout over all the eight systematic strains. On average, each of the three combinatorial gene knockouts provide an increase to each of the eight systematic backgrounds (boxes above 1 and small whiskers) while the impact of stacking these targets depends on the genetic background, but is often detrimental (boxes below 1, but with large whiskers reaching to values above 1).

5.3.2 Hierarchical Clustering Analysis

Clustering methods have been routinely applied for the analysis of microarray (and other) data to determine sets of genes that exhibit similar expression profiles (Eisen et al., 1998). Likewise, the technique of hierarchical clustering may be applied to the metabolic landscape of **Figure 5.5** in order to cluster gene knockout constructs exhibiting similar production profiles over the four time points. Presumably, strains clustering most closely accumulate product by following similar mode-of-action in the mechanism of lycopene production. To this end, we performed a complete linkage hierarchical clustering of the lycopene time profiles for the entire 8x8 strain matrix using the Euclidean distance as the similarity metric. Upon clustering the entire set of 64 strains, two distinct organizations emerge for the two sets of gene targets.

5.3.2.1 Stoichiometric targets have similar modes-of-action

Clustering lycopene profiles (across the four time points) for the eight stoichiometric knockout strains revealed a fairly close, stacked dendrogram (**Figure 5.7**). The lycopene levels for the strains at the four timepoints are presented by the heat plot. This stacking is in concert with the presumed mode-of-action in these strains, namely the increasing availability of precursors and cofactors that are needed for lycopene biosynthesis. This is further evidenced by the close clustering of strains like Δ*fdhF* and the parental strain, as the *fdhF* single knockout was determined from the stoichiometric analysis to bring about no enhancement of lycopene production.

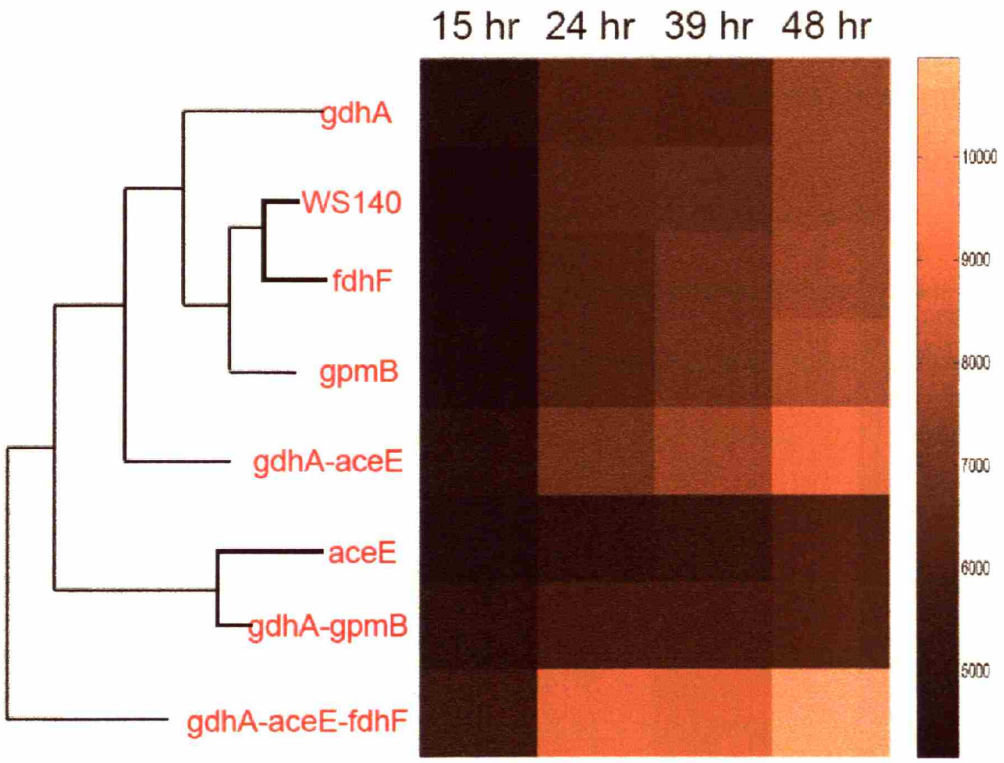

Figure 5.7: Clustering analysis of time course data for systematic targets.

Lycopene production profiles across the 48 hour shake-flask fermentation are clustered, resulting the in the dendrogram illustrated. The purely systematic strains have a stacked dendrogram which indicate a similar mode of action of lycopene production.

5.3.2.2 Combinatorial targets decouple modes of action of stoichiometric targets

The results presented by the box-and-whisker plot in **Figure 5.6** illustrate that the impact of stacked-combinatorial targets is different than single targets. These conclusions are further reinforced by a time-course clustering analysis depicted in **Figure 5.8**. Furthermore, the stoichiometric design implicit in the stacked dendrogram is altered through the addition of any combinatorial genotype. As an example, **Figure 5.9** shows the clustering of lycopene time profiles for the knockout strains obtained by combining each of the 7 stoichiometric targets with the combinatorial target gene *hnr*. In contrast to **Figure 5.7**, all combinatorial targets, as exemplified by *hnr,* force a split-tree shape in the dendrogram. Different time courses in lycopene accumulation suggest different modes-of-action for the effect of the combinatorial genes on this phenotype. Specifically, while each of the single knockout constructs formed from the combinatorial targets tend to exhibit similar behavior (increased production), the combination of these genes is not linear or synergistic. In fact, the double and triple knockout constructs arising from these combinatorial targets exhibit vastly different production profiles from the individual targets. This non-linearity suggests that the combinatorial targets are disrupting regulatory processes that are relatively incompatible, and in certain cases deleterious, when combined.

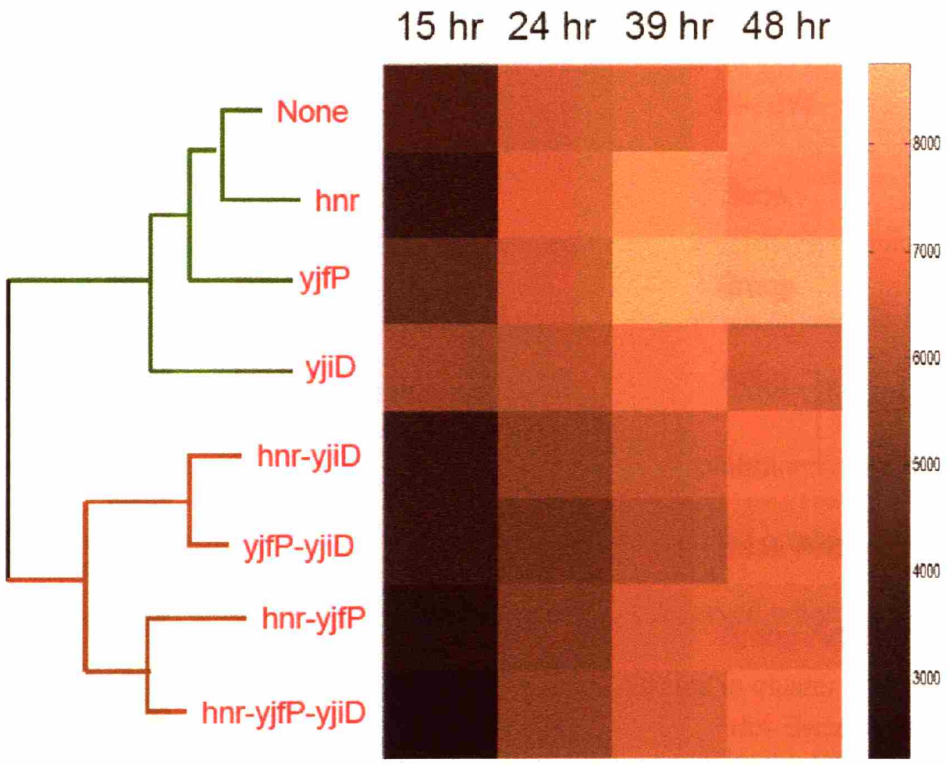

Figure 5.8: Clustering analysis of time course data for combinatorial targets. Lycopene production profiles across the 48 hour shake-flask fermentation are clustered, resulting the in the dendrogram illustrated. The single target combinatorial strains behave quite differently than those strains possessing multiple combinatorial target gene knockouts.

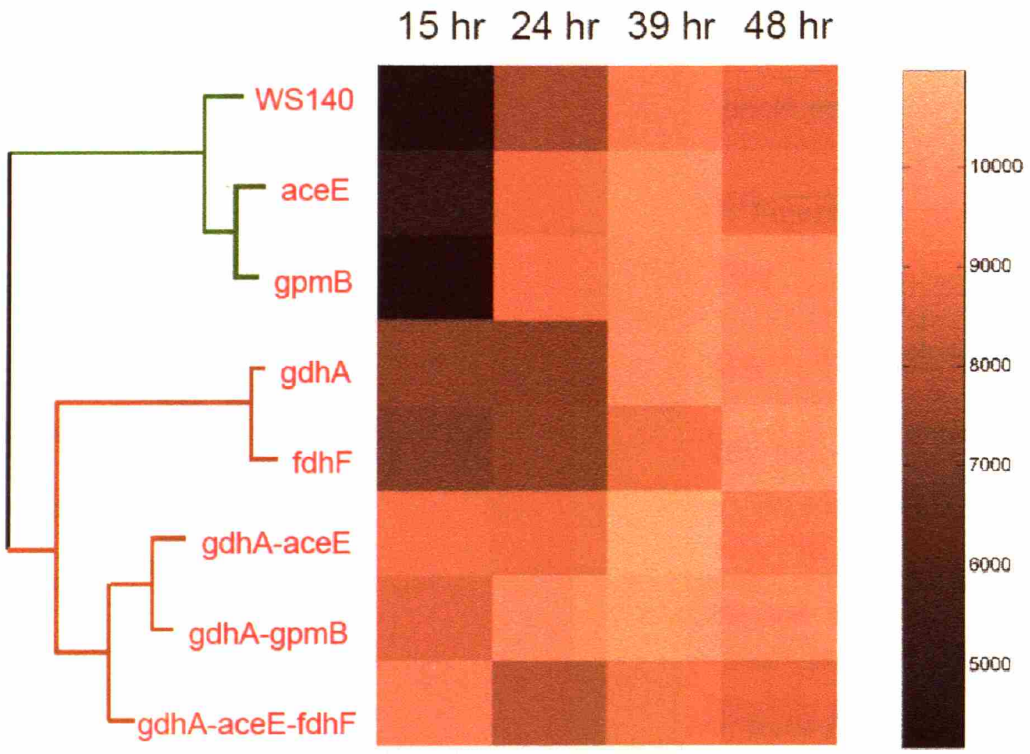

Figure 5.9: Clustering analysis of time course data for systematic targets with *hnr* knockouts. Lycopene production profiles across the 48 hour shake-flask fermentation are clustered, resulting in the dendrogram illustrated. The previously stacked dendrogram now exhibits a split-tree form where there is a clear separation in the production profiles between the various genotypes. The addition of any combinatorial genotype (as illustrated here with *hnr*) results in this decoupling of the systematic design.

5.3.2.3 Clustering analysis highlights varied modes-of-action

Figure 5.10 presents the results of the clustering analysis for the systematic strains in the absence (**Figure 5.10A**) and presence (**Figure 5.10B**) of an *hnr* knockout. When the systematic strains are plotted against the lycopene accumulation level, they reveal an expanding concentric bubble-plot suggesting an *additive* effect of accumulating gene deletions. However, differences are observed when combinatorial genes are combined with stoichiometric ones. **Figure 5.10C** presents the production profiles of the three clusters of genotypes. Biological differences are observed when combinatorial genes are deleted together with stoichiometric ones. Strains in cluster *Y* all exhibit an extended lag phase which extends to 16-18 hours before reaching a typical cell density OD 3.5 – 4.0. In contrast, strains in cluster *Z* do not posses such a lag phase and exhibit a steady increase of lycopene production with time. The average, scaled production profiles for the purely systematic cluster and the two clusters forced by an *hnr* deletion are compared in **Figure 5.10C**. It is noted that this branched pattern is exhibited by all strains constructed from the deletion of any combinatorial gene in the background of the stoichiometric targets, with different production profiles characterizing each of the clusters.

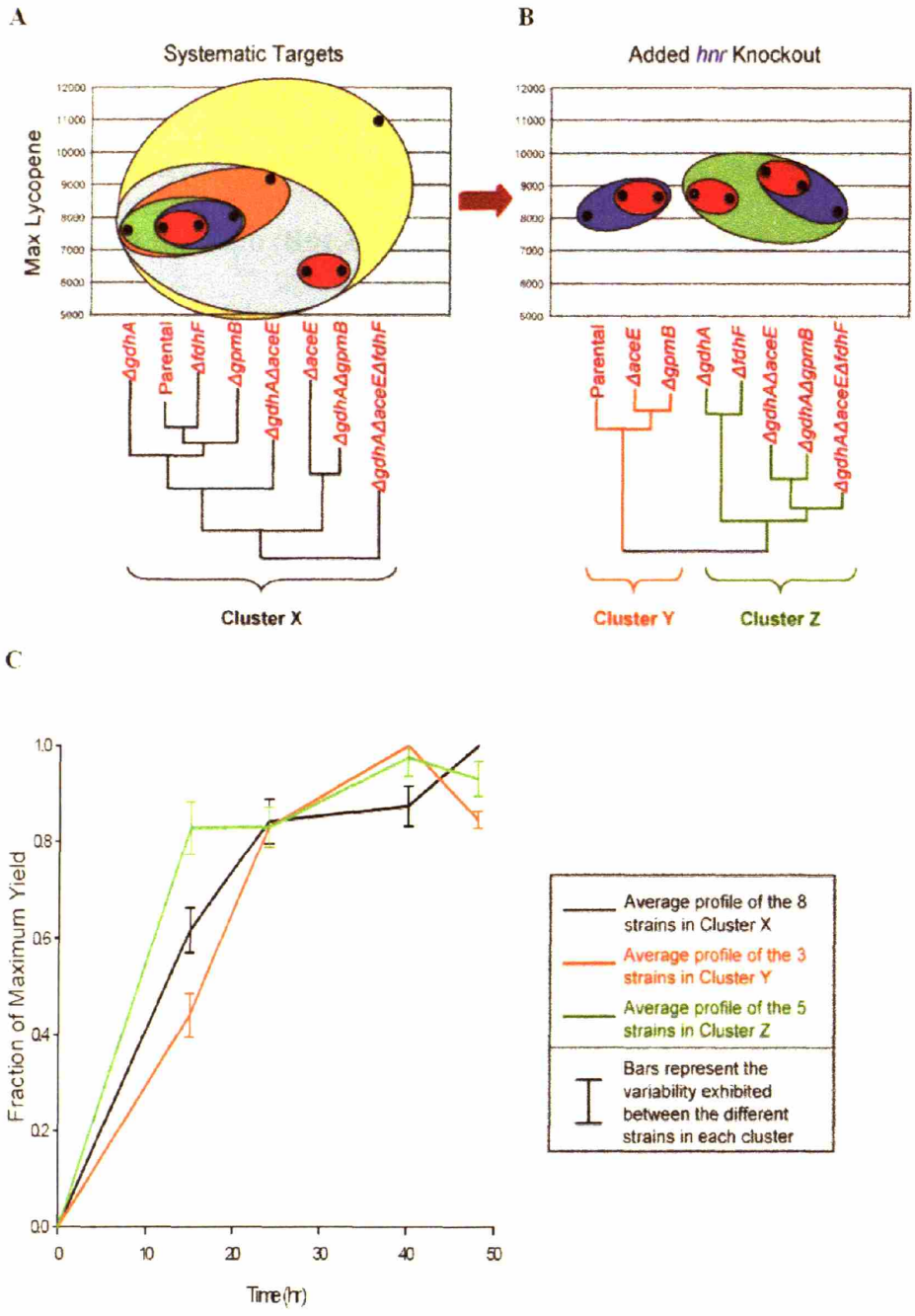

Figure 5.10: Clustering analysis and bubble plots. A summary of clustering analysis and bubble plots illustrates that local, metabolic gene targets are more accessible through a sequential search than global, regulatory targets which require a simultaneous search which is sensitive to the genetic background of the strain. Panel **C** compares the average, relative production profiles for the three clusters shown in Panels **A** and **B**.

5.3.3 Covariance analysis

As a reinforcement of the results described above, statistical metrics may be used to quantify the interaction between systematic and combinatorial gene targets. As such, quantitative metrics beyond clustering can asses the source of non-linearity in the production phenotypes of the 64 strains. **Figure 5.11** and **Figure 5.12** present the results of a covariance analysis between the 64 strains in this collection. Covariance analysis in this context essentially quantifies, in a pair-wise fashion, the level of association between the lycopene production levels. The covariance analysis quantifies the correlation between the values in two different 8 x 1 matrices. First, covariance analysis of strains with a given systematic target (1 systematic x 8 combinatorial) yields all positive values with many of the various backgrounds exhibiting a high covariance and thus correlation (**Figure 5.11**). Conversely, strains with a given combinatorial genotype (8 systematic x 1 combinatorial) had both positive and negative covariances with other genotypes which illustrates the lack of similar modes of action between the combinatorial targets and within the different systematic targets (**Figure 5.12**). As a result of these analyses, the major nonlinearities entering into the genotype-phenotype landscape are mostly due to regulatory or unknown factors as opposed to stoichiometry.

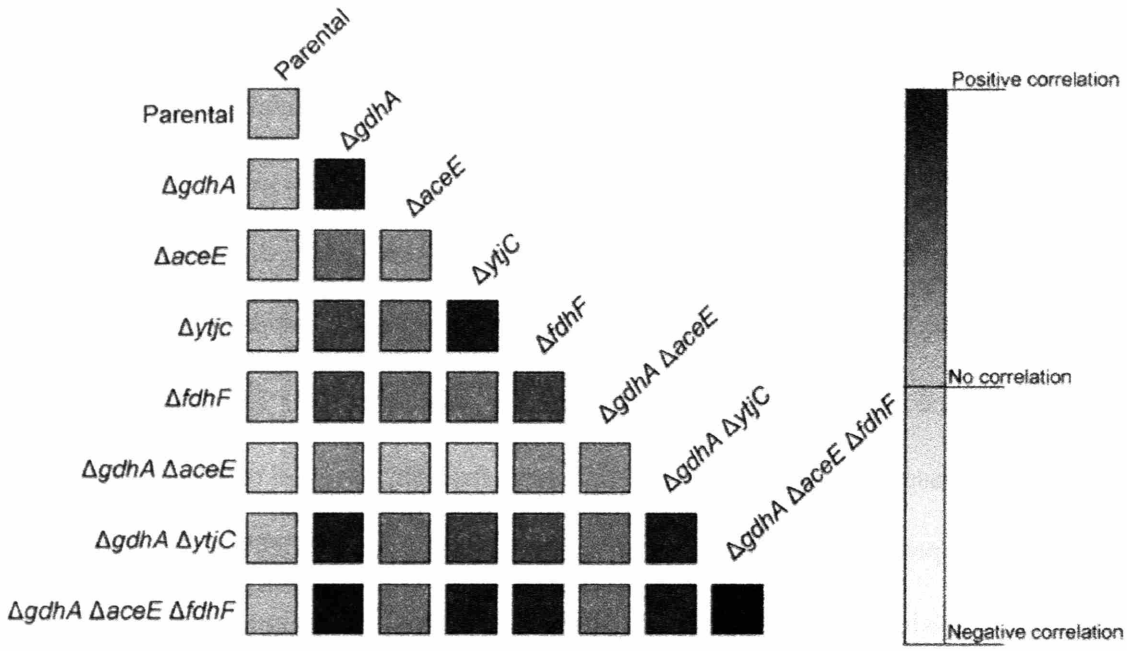

Figure 5.11: Covariance analysis of systematic targets. A covariance analysis on the lycopene production values was performed between the 64 strains in this collection. In this context, values for the covariance were calculated between the values in two different 8 x 1 matrices. The intensity of the square is proportional to the value of the covariance and is qualitatively represented in the scale with the midpoint representing a covariance of zero. Covariance analysis of the strains across the systematic genotypes (1 systematic x 8 combinatorial matrix) yields all positive values with many of the various backgrounds exhibiting a high covariance and thus correlation. In this type of analysis, each of the squares represents the covariance of all strains (8 in total) containing the systematic genotype of the column with that of the all the strains (8 in total) containing the systematic genotype of the row.

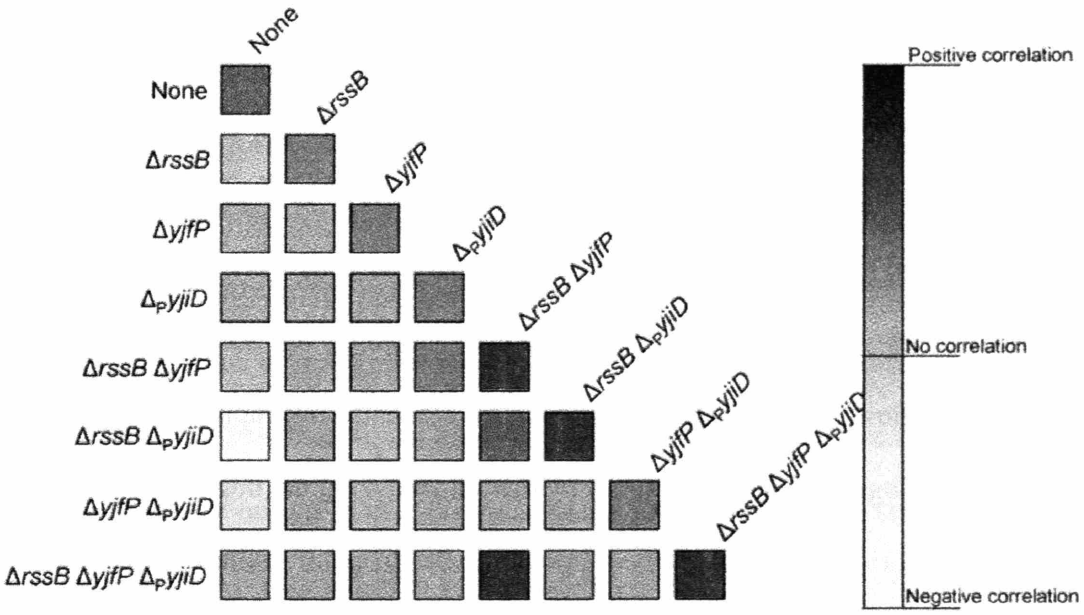

Figure 5.12: Covariance analysis of combinatorial targets. A covariance analysis on the lycopene production values was performed between the 64 strains in this collection. In this context, values for the covariance were calculated between the values in two different 8 x 1 matrices. The intensity of the square is proportional to the value of the covariance and is qualitatively represented in the scale with the midpoint representing a covariance of zero. Covariance analysis of the strains across the combinatorial genotypes (8 systematic x 1 combinatorial matrix) yields both positive and negative values, with most genotypes showing no significant covariance. These results illustrate that the combinatorial targets introduced the major source of non-linearity in this collection of strains. In this type of analysis, each of the squares represents the covariance of all strains (8 in total) containing the combinatorial genotype of the column with that of the all the strains (8 in total) containing the combinatorial genotype of the row.

5.3.4 Summary

As illustrated by these different analyses, combining stoichiometric and combinatorial targets creates a complex metabolic landscape with several local and global optima. The nonlinear effects of the regulatory and unknown targets lead to the complex topology of this landscape. In particular, stacking combinatorial targets upon the systematic design of the stoichiometric targets leads to a decoupling of the stoichiometric logic. This decoupling is evident in analyzing the impact of the *hnr* or any other combinatorial constructs on the shape of the dendrogram resulting from hierarchical clustering of the time-series data for the eight stoichiometric strains. These results suggest the capacity to search for metabolic or local targets through an additive, sequential search for gene targets, such as the one undertaken here. On the other hand, kinetic, regulatory and other unknown factors generate strong non-linear effects requiring a simultaneous search approach. While identification of optimal gene targets will continue to be a demanding undertaking, searches for gene targets will be significantly aided by advanced models of cell function accounting for kinetic and regulatory mechanisms.

5.4 Optimizing fermentation profiles

The exhaustive exploration of the combinations of stoichiometric and combinatorial targets allowed the identification of several strains of interest on the basis of their performance in small, batch shake-flask cultivations. To better assess the production capacity of these knockout strains, fed-batch cultivations were carried out in

shake-flasks and bioreactors under controlled conditions with staged glucose feed. Several strains corresponding to interesting optimum points in the landscape of **Figure 5.5** were selected for further characterization. **Figure 5.13** presents the results of optimized shake-flask fermentations. These results illustrate the capability of the global maximum strains to produce upwards of 18,000 ppm in 24 – 40 hours. Of course, bioreactor optimization is an iterative process and it is possible to increase yields through improving control. Nevertheless, the high correspondence between these more optimized fermentations and the original shake-flasks suggests that the behavior is similar and thus the general trends are conserved even with this change of culturing environment. More importantly, while initial shake-flask fermentations in glucose minimal media showed that these two global maxima strains behaved similarly, with each producing 11,000 PPM (μg lycopene/g DCW) over the course of a 48 hour fermentation, a more optimized staged glucose feed experiment suggested that the Δ*gdhA* Δ*aceE* Δ$_{P}$*yjiD* strain had a higher rate of production reaching 18,000 PPM in just 24 hours (compared with 40 hours for the Δ*gdhA* Δ*aceE* Δ*fdhF* strain). Therefore, it appears that the performance of these mutants is dependent on the culturing conditions, and possibly other environmental factors which can be optimized in a bioreactor setting.

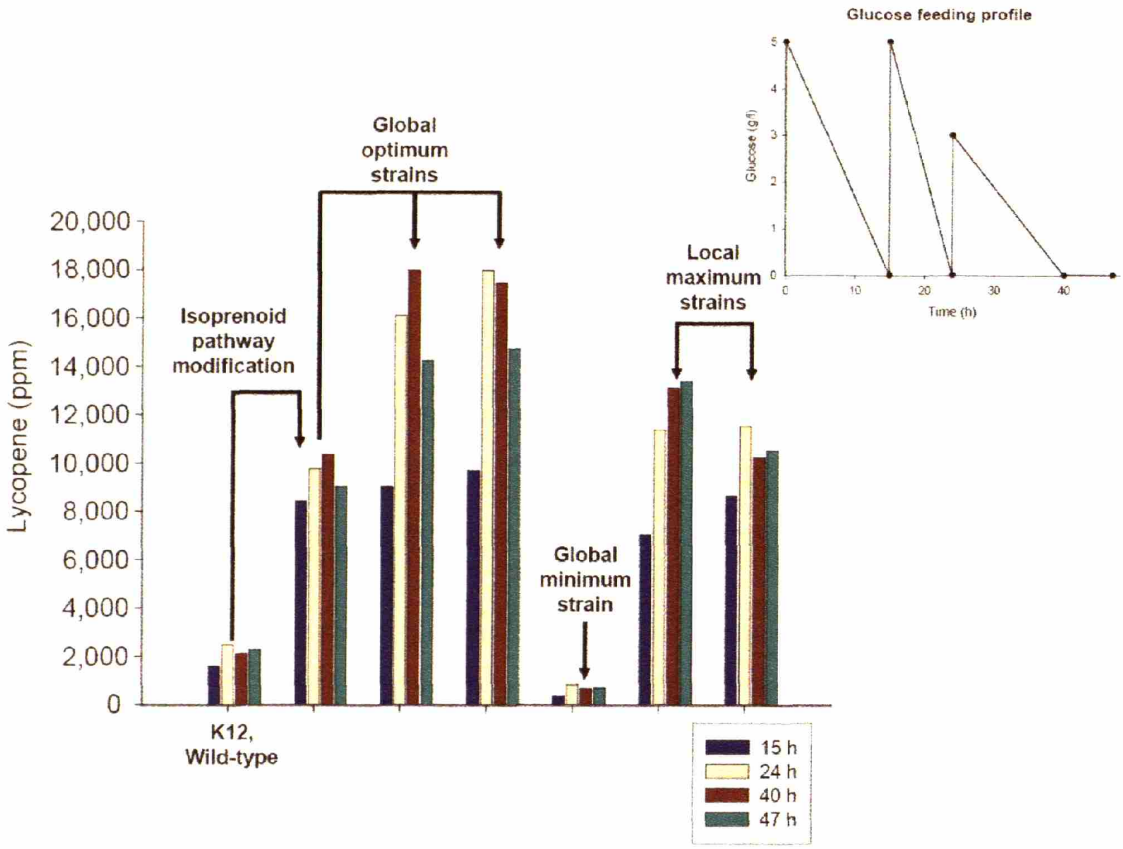

Figure 5.13: Behavior of selected strains in optimized culturing conditions.

Selected strains from the metabolic landscape were cultured in fed-batch shaker flasks with increased M9 salts and a staged glucose feed as represented in the graph. Strains presented from left to right are (1) K12 (wild-type), (2) engineering parental strain with *dxs*, *idi*, and *ispFD* overexpressions, (3) Δ*gdhA* Δ*aceE* Δ*fdhF*, (4) Δ*gdhA* Δ*aceE* Δ$_P$*yjiD*, (5) Δ*gdhA* Δ*aceE* Δ*fdhF* Δ*hnr* Δ*yjfP*, (6) Δ*gdhA* Δ*aceE* Δ*fdhF* Δ*yjfP*, (7) Δ*gdhA* Δ*aceE* Δ*hnr* Δ*yjfP* Δ$_P$*yjiD*. The two global maxima were capable of producing upwards of 18,000 ppm in 24 to 40 hours.

5.5 High cell density fermentations

Previous work (Alper, Miyaoku, & Stephanopoulos, 2005) indicated that dry cell weight (dcw) specific lycopene productivity (as measured in μg of lycopene per cell mass) could be increased by altering salt concentrations and optimizing glucose feeding profiles. A 2x M9 medium was found to be optimal for our purposes (Alper, Miyaoku, & Stephanopoulos, 2005).

5.5.1 Determination of optimal fermentation parameters

Before optimizing the glucose feed, we sought out to investigate how additional key fermentation control parameters such as agitation speed and pH influence the production profiles of lycopene. All reactors were set at 37°C, which is the optimal temperature for cell growth.

5.5.1.1 Agitation

Agitation speed was investigated as a putative bioreactor parameter responsible for controlling dissolved oxygen content and maximum cell density. Stoichiometric analysis (Alper et al., 2005b) suggested that volumetric lycopene yield in *E. coli* from glucose increases with oxygen uptake rate, due to the large energetic requirement of lycopene production (8 CTPs and 8 ATPs per mole). We carried out a series of cultures at various agitation speeds and **Figure 5.14** presents the volumetric production level of lycopene after 24 hours as a function of the sampled agitation speeds. This timepoint was chosen as it represents the time at which glucose had been fully utilized from the last

pulsed feed. The increasing trend and higher volumetric productivities at increased agitation speeds matched well qualitatively with the findings of stoichiometric analysis. These results indicated that the lycopene fermentations should be run under aerobic conditions during times of balanced growth and glucose utilization.

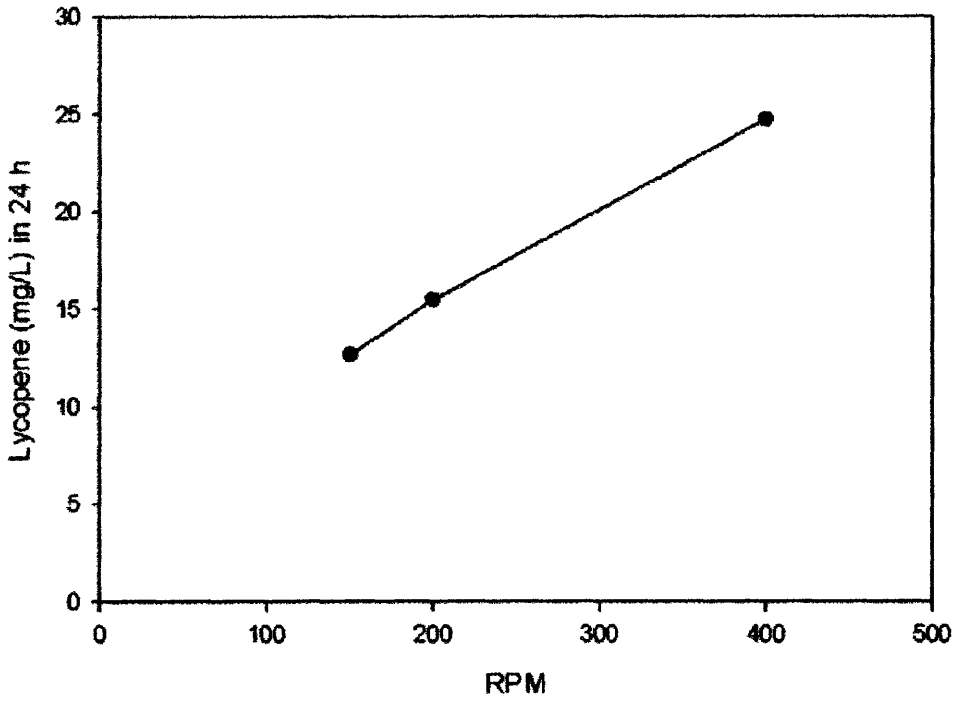

Figure 5.14: Fermentation-based investigation of oxygen level on lycopene. Trial batch fermentations for the Δ*gdhA* Δ*aceE* Δ*pyjiD* strain in 2x M9 media with a staged glucose feed show an increasing trend of volumetric lycopene production as a function of agitation, an operating parameter used to control aeration. These results qualitatively match the *in silico* predictions.

5.5.1.2 pH control

The control of pH in a bioreactor can influence cell growth and consequently, secondary metabolite production. Furthermore, pH changes in a reactor could introduce adverse effects on the growth rate of the culture. Varying strategies of double-side and partial (single-side) pH control were tested with respect to lycopene production. As an illustration of the pH effect, a fully controlled pH 7.0 strategy was compared with a partial (only base) control strategy. **Figure 5.15** depicts the time profile of lycopene production under a constant pH 7.0 controlled through the addition of NH_4OH and HCl. While the lycopene level remains constant for times after 24 hours, the cell density is still increasing. As a result, lycopene is being produced at the same rate at which it is being diluted by growth. Conversely, the profile of specific lycopene productivity under partial pH control continues to increase after 24 hours and is accompanied by a significant pH increase in the stationary/production phase (**Figure 5.16).** This phenomenon is supporting evidence that reduced growth rate results in increased specific productivity levels. The observed pH increase correlated with an increased specific lycopene productivity. Under this control strategy, the total cell density did not increase appreciably after 24 hours (**Figure 5.16**). Therefore, the results suggest that a partial pH control with only base addition allowed with a starting pH of 7.0 had the best production profile of lycopene.

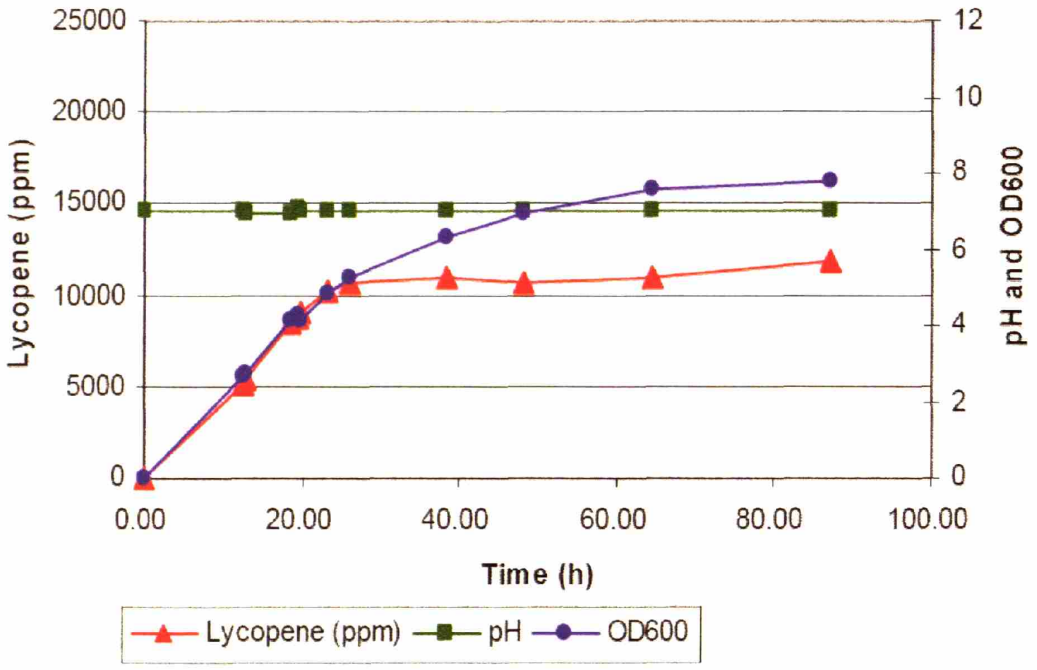

Figure 5.15: Impact of double-sided pH control on lycopene production. pH control strategies were varied for ΔgdhA ΔaceE $\Delta_P$yjiD cultures in a 2x M9 media batch fermentation using a staged glucose feed. For the case of a constant pH of 7.0 controlled with both acid and base, the specific lycopene production (in PPM) plateaus after around 24 hours. However, cell density continues to increase after this point and therefore, lycopene production increases at the same rate at which it is being diluted by growth.

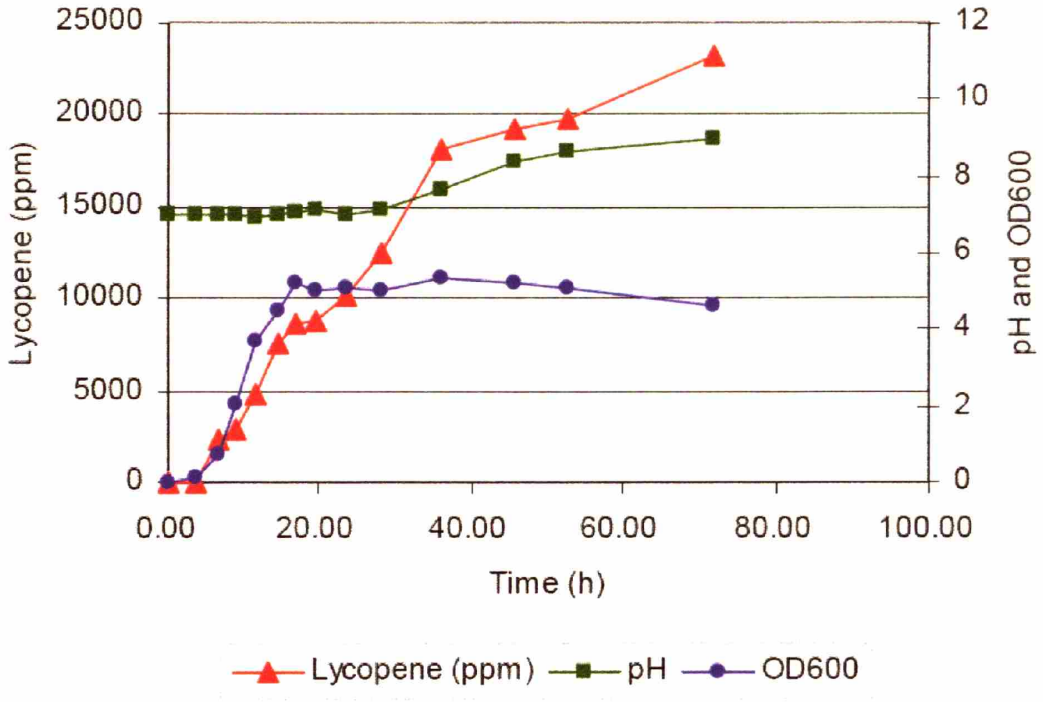

Figure 5.16: Impact of single-sided pH control on lycopene production. pH control strategies were varied for ΔgdhA ΔaceE $\Delta_P$yjiD cultures in a 2x M9 media batch fermentation using a staged glucose feed. When the reactor is only controlled with base and allowed to be at or above pH 7.0, the lycopene production shows significant increases past the 24 hour timepoint which correlates to drastic increases in the pH level, however is not accompanied by a significant increase in cell density. Comparing this profile to that of double-sided control suggests that single-sided pH control is a more favorable condition for increasing specific yields.

5.5.2 High cell density fermentations

Since lycopene is stored as an intracellular product in the membrane, special considerations must be made for identifying optimal cultivation parameters. Optimal parameters must be able to promote lycopene production during both growth phase and stationary phase. During growth phase, a primary concern is the volumetric production (mg of lycopene per liter in the reactor per hour) and optimization of product yield (g lycopene per g glucose). Furthermore, it is important during this phase of the bioreactor to concurrently accumulate a high quantity of biomass. During stationary phase, the specific productivity (μg lycopene per g dry cell weight per hour) will increase. The identification of bioreactor conditions which encourage both modes of lycopene production is important in creating an optimized process. To accomplish this, many external parameters may be optimized to balance between these two modes of production. Specifically, the trade-off relationship between cell growth rate and specific lycopene productivity is an apparent limitation in obtaining high lycopene volumetric productivity, which is industrially an important process metric. In order to release this limitation, fed-batch fermentations were conducted that allowed high cell density to be obtained during the growth phase followed by production during the stationary phase.

Fermentations of the parental control strain along with the two engineered triple knockout strains were conducted using a controlled glucose feed with partial pH control and high aeration rates (by continually increasing the RPM). This feeding and control strategy was optimized based on the trial fermentations described above. The production profiles of lycopene in mg/L and PPM are presented in **Figure 5.17** and **Figure 5.18**

respectively. All strains exhibit a nearly similar growth associated volumetric lycopene production rate. However, significant differences between the engineered strains and the parental strain are seen after around 15-18 hour from the start of the reactor. Lycopene accumulations (both specific and volumetric) remain constant after 15 – 18 hours for the parental strain. However, significant increases in the volumetric productivity and specific productivity are seen in both engineered strains (**Table 5.1**). Furthermore, it is clear that the production profiles are indeed distinct for these two strains suggesting evidence that the different genotypes are leading to different production phenotypes. Fermentation metrics for the growth characteristics and lycopene production rates are summarized in **Table 5.1**. The two optimized strains exhibit increased productivities and decreased growth rates compared with the parental strain. Also, the engineered strains yielded more robust and reproducible fermentations as is exhibited by the smaller standard deviations in all categories (**Table 5.1**). The fermentor pH was controlled according to the one-sided control strategy using only base addition. Similar to the trial profile in **Figure 5.16**, the pH did increase during the production phase to reach values upwards of 8.0 at around the 24 hour timepoint.

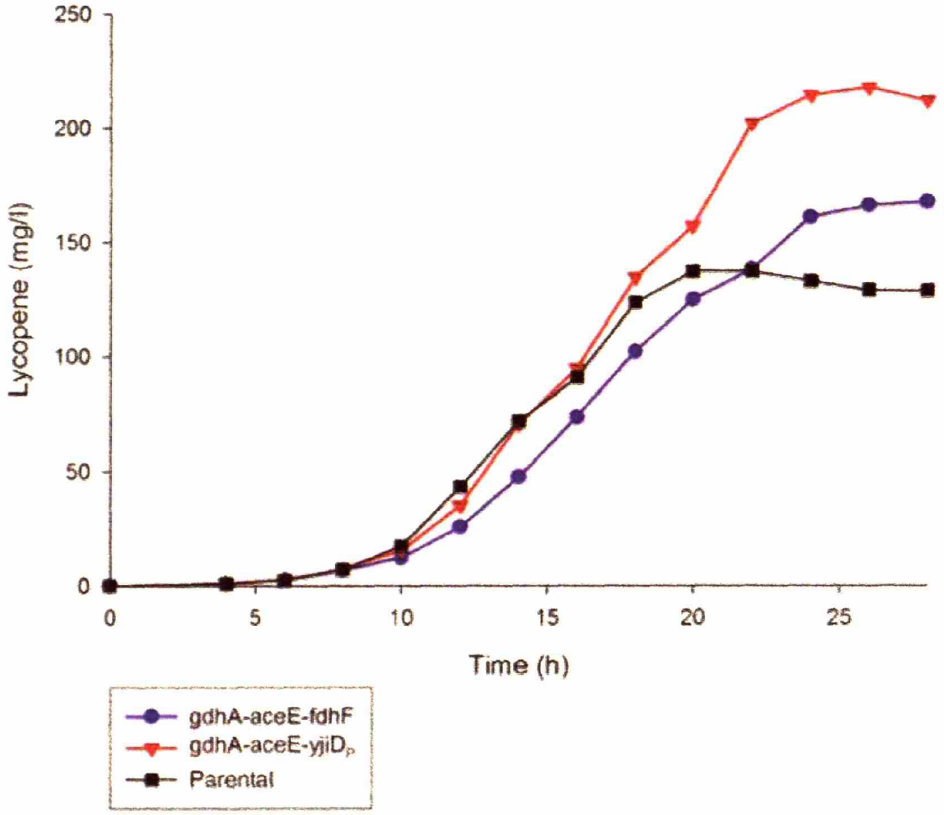

Figure 5.17: Volumetric lycopene production in the reactors. Lycopene (presented in mg/L) accumulates as a growth associated product and growth independent product. In the parental strain, lycopene levels remain nearly constant after 15-18 hours. However, the two engineered strains both exhibit significant increases in lycopene levels well after cell growth stops. Furthermore, the production profiles of the two engineered strains are distinct.

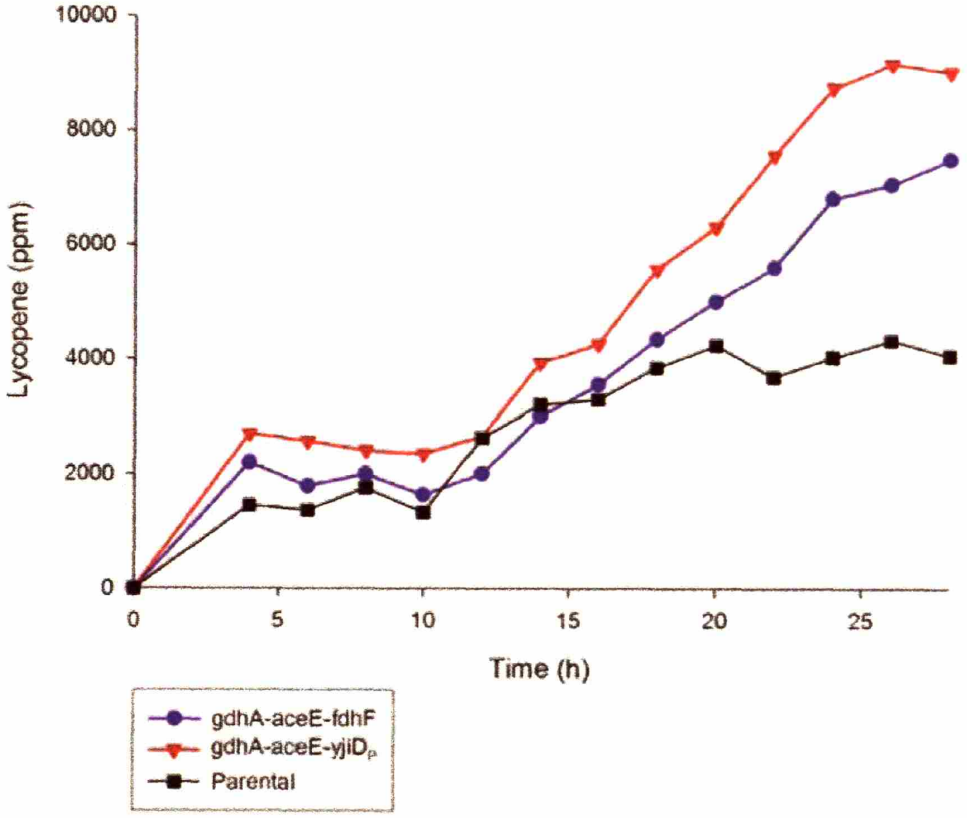

Figure 5.18: Specific lycopene production (ppm) in the reactors. The specific lycopene content in PPM (μg/gdcw hr). This graph illustrates the significant difference in the lycopene content per cell in the engineered strains compared with the parental strain. As with the volumetric productivity, accumulation continues for the engineered strains after the 15-18 hour timepoints. Average values for the replicate experiments are presented in both graphs.

		Parental	ΔgdhA ΔaceE ΔfdhF	ΔgdhA ΔaceE Δ_{P}yjiD
Growth Phenotype	μ (hr^{-1})	0.504 (.003)	0.408 (.016)	0.446 (.019)
	Maximum OD_{600}	92.1 (4.4)	89.5 (3.8)	84.6 (1.7)
Lycopene Phenotype	Specific Production (μg/gdcw hr)	242.9 (70.1)	388.0 (30.6)	496.3 (15.6)
	Bioreactor Productivity (mg/bioreactor hr)	12.3 (2.8)	13.7 (0.5)	17.9 (0.6)
	Total Produced (mg) (24 hr)	139.1 (39.4)	186.2 (20.4)	228.5 (10.6)
	Maximum mg/L	132.7 (38.3)	176.5 (12.8)	221.6 (7.3)
	Overall Yield of reactor (mg/g glucose) (24 hr)	1.58 (0.43)	2.15 (0.26)	2.61 (0.11)

Table 5.1: Growth and lycopene phenotypes of strains in fed-batch reactor. The average growth rates and lycopene production rates are presented from duplicate high cell density fermentations. Numbers in parenthesis represent the standard deviation of these values from two separate trials for each strain. The two optimized strains exhibit increased production rates and decreased growth rates compared with the parental strain. Also, the engineered strains were more stable and reproducible as is exhibited by the smaller standard deviations in all categories.

5.5.3 Carbon balances

Samples were taken from the bioreactors in two hour intervals to measure organic and amino acid levels, as well as the lycopene level and biomass. Total carbon balances are created by exhaustively measuring the extracellular carbon-based molecules in an effort to account for all glucose used. These balances (**Figure 5.19**) suggest similar distributions of carbon among the products of all the strains with a slight increase in carbon dioxide contribution for the engineered strains. However, a more detailed analysis of four components (lycopene, formate, glutamate, and alanine) supplying marginal contributions (collectively <3%) to the carbon balance (**Figure 5.20**) reveals significant differences in the carbon fluxes. A significant amount of glutamate was detected in the media for the parental strain when compared to the secreted level of the two engineered strains, owing to the deletion of the *gdhA* gene in these two strains. Furthermore, lycopene production was inversely related to the level of glutamate secreted by the various strains. Stoichiometric analysis also previously indicated that a reduction in the glutamate flux (especially through *gdhA*) is related to an increase in lycopene production, presumably through the resulting increase in NADPH availability (Alper et al., 2005b). Furthermore, formate levels were substantially higher in the Δ*gdhA* Δ*aceE* Δ*fdhF* mutant, which is designed to limit the loss of carbon (particularly pyruvate) through formate and its subsequent byproducts. Interestingly, the other engineered strain (Δ*gdhA*, Δ*aceE*, Δ_P *yjiD*) showed a substantial decrease in both formate and glutamate levels, perhaps providing some clues to the function of the hypothetical protein encoded by *yjiD*. Moreover, alanine was also detected in the media and despite its small

contribution, was found to contribute five fold less in the Δ*gdhA*, Δ*aceE*, Δ_P *yjiD* strain than the other two strains, which were nearly similar.

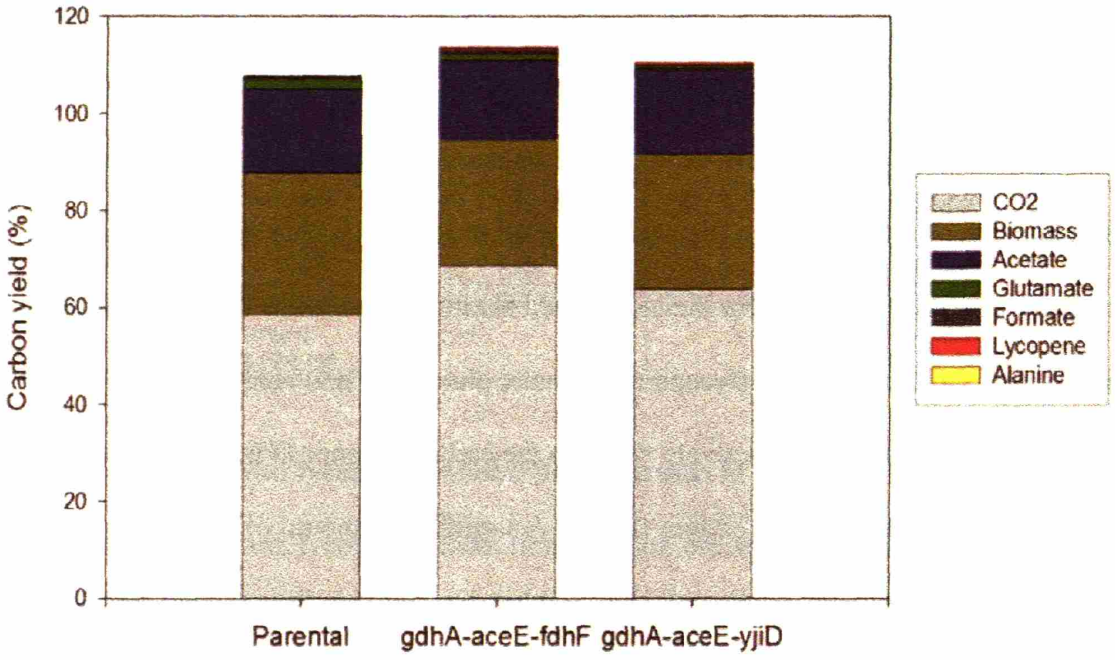

Figure 5.19: Overall carbon yield balances for the fermentors. Organic and amino acids, biomass and lycopene were measured off-line, while carbon dioxide was measured online. The average final carbon yields are presented for each of the strains. Most of the glucose fed to the reactor went to carbon dioxide and biomass. The contribution of the four components accounting for the smallest amount of the carbon (glutamate, formate, lycopene and alanine) are also depicted in **Figure 5.20**.

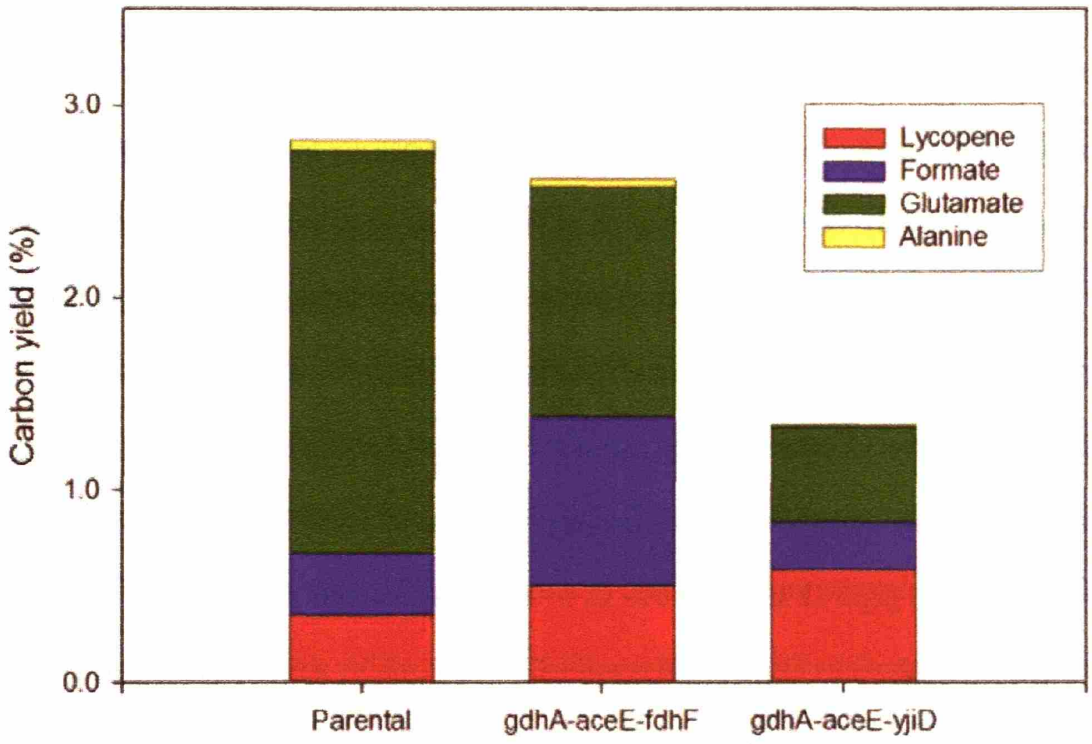

Figure 5.20: Marginal carbon yield balances for fermentors. A more detailed analysis of four minor (or marginal) contributors to the carbon balance indicates a significant glutamate level in the parental strain. Furthermore, all three side-components (glutamate, formate, and alanine) were reduced in the Δ*gdhA* Δ*aceE* Δ*pyjiD* triple mutant. These differences in marginal contributors, while significant, are hard to decifer from viewing **Figure 5.19** as they only account for less than 3% of the total carbon balance.

5.5.4 Bioreactor summary

These studies confirmed that the production profiles of the two engineered strains are indeed distinct suggesting that the different genotypes are leading to different production phenotypes. Further analysis of the secreted metabolite levels was able to suggest some of the underlying factors related to lycopene production including the importance of glutamate, alanine and NADPH flux. Recently, we have analyzed these strains at the level of gene transcription and metabolome profiles in an effort to gain a further understanding of the mode of action of these gene knockouts. These preliminary results suggest that the true phenotype of these strains must be assessed at multiple dimensions rather than strictly by the production level of lycopene. Furthermore, rational design using the collected metabolomic and genomic data and additional explorations of the metabolic landscape can further identify gene knockout or over-expression targets which may optimize these strains further and increase lycopene yields. These include identifying alternative pathways for increasing NADPH availability as well as decreasing the secreted glutamate further as this value negatively correlates with accumulated lycopene concentration. Nevertheless, by the end of this study, the optimized high cell density fermentations of these engineered *E. coli* strains resulted in titers of around 220 mg/L. These titers are the highest reported in literature to date for the production of 40 carbon carotenoids in *E. coli*.

5.6 Summary

The identification of multiple gene targets which impart material improvement of a particular phenotype is an open problem. Among the complications are strong non-linear effects, lack of accurate models capable of capturing genetic interactions, and ineffective search strategies. This chapter addressed an exhaustive experimental search to investigate the effect of combining rationally selected genes with those identified through combinatorial methods. The resulting search yielded a number of promising strains some of which were capable of producing upwards of 22,000 PPM (or 22 mg/g DCW) in defined glucose media when cultured under fed-batch conditions. This value represents a nearly 4-fold increase over the parental strain when cultured in simple cultivation and a 2 fold-increase over the pre-engineered parental strain in similar conditions. Furthermore, this represents an 8.5-fold increase to the wild-type K12 *E. coli* strain when cultivated in these similar conditions. Subsequently, these strains were cultivated in optimized fed-batch bioreactors which resulted in titers of around 220 mg/L which are higher than values reported in literature.

The metabolic landscape defined through this unique set of 64 knockout strains provides the basis for several observations of importance to metabolic engineering. First, rationally selected stoichiometric gene knockout targets have the potential of generating serious contenders in the quest of maximally producing strains. It is noted that one of the two maximum overproducing strains resulted from the knockout of three stoichiometric genes (*gdhA, aceE, fdhF*). In addition, the knockout of specific combinatorial genes

yielded significantly enhanced phenotype in the background of particular stoichiometric knockout genes. Second, while combinatorial gene targets hold greater potential than stoichiometric ones as *single* knockout mutants, multiple knockouts of the combinatorial gene set led to a distinct deterioration of the lycopene phenotype. On the other hand, the second global maximum could be obtained only from a combination of rational and combinatorial gene targets. These observations suggest that, while the effects of multiple knockouts may be additive and more predictable for stoichiometric genes, strong non-linear effects characterize the function(s) of the combinatorial genes. Thus it appears that rational gene target selection through stoichiometric modeling identifies several metabolic targets with similar mode-of-action that can be translated across different genetic backgrounds, including those of single knockouts of combinatorial genes. Third, the presence of many local maxima complicates the nature of the landscape and raises questions about general sequential search strategies. By exhaustively evaluating computationally all pair-wise knockout combinations using a genome-wide stoichiometric model for *E. coli*, sequential search strategies were found to be quite effective when applied to the space of stoichiometric genes (Alper et al., 2005b). **Figure 5.10** suggests that this result does not hold when combinatorial genes are also included in the search space, necessitating exhaustive combinatorial searches of the type undertaken in this study.

It should be noted that the search of this study was limited to the effect of gene knockout only. No gene knockdown or over-expression was considered. While these options of genetic modulation add an extra layer of complexity in the metabolic engineering of overproducing strains, they hold nevertheless vast potential in the

effective redistribution of cellular metabolic processes for further drastic improvements of product over-production phenotypes.

Efficient phenotype optimization necessitates a robust, defined search strategy to identify genetic targets requiring modification. While non-linearity and multiple-optima make the generalization of such a strategy difficult, this study elucidates several considerations of importance for optimizing phenotype. First, the development of high-throughput screening combined with detailed cellular models aids efficient strain optimization. Second, combinatorial targets influencing global cellular function should be identified at later stages in the strain improvement process to avoid selecting those with limited utility or incompatible modes-of-action. Finally, these results suggest that metabolic genes have a linear impact in the overall cellular phenotype while regulatory or unknown targets have more non-linear impacts on cellular function. This study serves as the first case-study for understanding the complex interaction of the genotype-phenotype space in the context of a product over-production phenotype. In this case, the exhaustive exploration of combinations of gene knockout targets arising from the application of systematic and combinatorial methods yielded lycopene overproducing strains. As such, the lessons gained here can help shape future strain improvement programs as they are tested in diverse systems for divergent products.

At the beginning of this chapter, it was indicated that two major outstanding questions exist about metabolic landscapes: (1) how non-linear are metabolic landscapes and (2) what is the optimal search strategy for traversing these landscapes. The results presented here focused on mapping the topology of the metabolic landscape. As such, it was found that high levels of non-linearity are introduced into the metabolic landscape,

especially by combinatorial knockout targets. Furthermore, it was shown that gene targets which were individually selected are not always universal targets, and the effectiveness of these targets depends on the specific genotype of the strain of interest. Ultimately, it is important to understand this non-linearity in an effort to create an optimal search strategy for traversing this landscape. Chapter 6 addresses the issue of search strategy through the further probing of this metabolic landscape through the identification of additional gene knockout targets using varying starting points in this landscape.

Chapter 6

The lycopene gene knockout search network

Systematic and combinatorial tools for the identification of gene knockouts were demonstrated to be effective in the improvement of cellular phenotype in the prior two chapters. It was previously demonstrated how these tools can be combined to identify strains of interest spanning the metabolic landscape (Alper et al., 2005b; Alper, Miyaoku, & Stephanopoulos, 2005). These investigations and subsequent analysis helped to shed light on the sources of non-linearity in the metabolic landscape and provided broad strategies for dealing with two distinct sets of gene knockouts (stoichiometric and regulatory/unknown). However, it is unknown how the overall search trajectory biases the exploration of the metabolic landscape. In particular, non-linearities in the metabolic landscape and the instance of recurrence in metabolic phenotypes confound the search for global maxima. In this chapter an iterative application of combinatorial gene knockout searches in *Escherichia coli* is conducted in the background of several lycopene overproducing strains to determine how the search is biased by the starting genotype. In particular, these combinatorial tools are employed in the background of eight different genotypes spanning various regions of the originally explored metabolic landscape.

Several interesting observations arise from this exploration of different gene knockout search trajectories. Initial examination of clusters and recurrence of gene knockout targets suggests key areas of metabolism correlating with lycopene productions. However, divergent genotypes indicate the potential of multiple, distinct paths to obtain comparable metabolic phenotypes. These targets and search trajectories are analyzed for their production potential and underlying mechanism. Ultimately, this chapter represents a culmination of the investigation of the metabolic landscape which started in Chapter 4. Through this more exhaustive search using different trajectories, we wish to address the following questions: (1) How do we successfully traverse complex metabolic landscapes, (2) Is there a unique mapping (one-to-one) of genotype to phenotype and (3) How can we extract the genotype-phenotype relationship most efficiently.

6.1 Probing the metabolic landscape

Combinatorial targets through the use of random transposon mutagenesis were identified in the background of eight strains: parental (pre-engineered strain with chromosome-based overexpressions in *dxs*, *idi*, and *ispFD*), ΔgdhA ΔaceE, ΔgdhA ΔaceE ΔfdhF, ΔgdhA, ΔyjfP, Δhnr, $\Delta_P$yjiD, and the ΔgdhA ΔaceE $\Delta_P$yjiD strain. Collectively, 800,000 mutants were analyzed (100,000 per each background) and 290 were selected for further characterization. **Tables 6.1-6.8** list the significant identified gene targets in each background and the annotated function. The inclusion of a "P" in front of a given gene name indicates that the transposon event occurred in the promoter region of the identified gene, similar to the *yjiD* promoter knockout discussed in Chapter

5. Each of these gene targets increases lycopene production (ppm) in the identified background to varying degrees ranging from 1.05-fold to 2.55-fold in 1x M9 minimal medium with 5 g/L of glucose. **Table 6.9** presents the average fold increase of all the selected targets in 1x M9 medium at the 15 and 24 hour timepoints relative to the respective controls.

hnr	σS degradation
P-*yjiD*	Hypothetical protein
yjfP	Hypothetical protein

Table 6.1: Identified gene knockouts in the parental strain background.

hnr	σS degradation
P-*yjiD*	Hypothetical protein
fdhA	(selB) selenocysteine incorporation (into *fdhF*)
yagR	Putative molybdemum cofactor-binding oxidoreductase
glnE	Protein adenylyltransferase, modifies glutamine synthase
pst	High affinity phosphate transport (membrane proteins)

Table 6.2: Identified gene knockouts in the Δ*gdhA* Δ*aceE* background.

hnr	σS degradation
fdhA	(selB) selenocysteine incorporation (into *fdhF*)

Table 6.3: Identified gene knockouts in the Δ*gdhA* Δ*aceE* Δ*fdhF* background.

hnr	σS degradation
yjfP	Hypothetical protein
lipB	Lipoate biosynthesis (related with *aceE* activity)
fdhA	(selB) selenocysteine incorporation (into *fdhF*)
clpXP	ATP dependent protease (one target: σS)
ygjP	Putative transcriptional regulator
yagR	Putative molybdemum cofactor-binding oxidoreductase
gntK	Gluconokinase
glnE	Protein adenylyltransferase, modifies glutamine synthase
modA	Periplasmic molybdate binding protein
ackA	Acetate kinase A

Table 6.4: Identified gene knockouts in the Δ*gdhA* background.

clpXP	ATP dependent protease (one target: σ^S)
glnE	Protein adenylyltransferase, modifies glutamine synthase
fdhD	*fdhF* Formation protein
fdhA	(selB) selenocysteine incorporation (into *fdhF*)
cyaA	Adenylate cyclase
aspC	Aspartate aminotransferase

Table 6.5: Identified gene knockouts in the Δ*yjfP* background.

fdhA	(selB) selenocysteine incorporation (into *fdhF*)
cyaA	Adenylate cyclase
yliE	Hypothetical Protein
sohA	Putative protease
pitA	Low affinity Phosphate Transport
yjhH	Putative enzyme
yfcC	Putative integral membrane protein
lysU	Lysine-tRNA ligase
yedN	Hypothetical Protein
P-yebB	Hypothetical Protein
ydeN	Putative enzyme (possible sulfur metabolism)
P-ycfZ	Putative Factor

Table 6.6: Identified gene knockouts in the Δ*hnr* background.

clpXP	ATP dependent protease (one target: σS)
ybaS	Putative glutaminase
P-appY	Acid (poly)phosphatase, starvation response
glxR	Tartronate semialdehyde reductase

Table 6.7: Identified gene knockouts in the Δ_P*yjiD* background.

moeA	molybdopterin biosynthesis
ackA	Acetate kinase A
nadA	quinolinate synthetase A
stpA	Putative regulator/chaperone
pstC	High affinity phosphate transport

Table 6.8: Identified gene knockouts in the Δ*gdhA* Δ*aceE* Δ_P*yjiD* background.

Table 6.9 (continued on next page)

Parental Strain		
	15 h	24 h
hnr	1.37	1.03
p-yjiD	1.31	0.98
yjfP	1.19	1.01
Δ*gdhA* Δ*aceE* background		
	15 h	24 h
fdhA	1.40	1.02
hnr	1.37	1.02
pst	1.29	1.33
yagR	1.18	0.97
nuoC	1.18	0.84
glnE	1.15	0.95
pta	0.97	1.10
Δ*gdhA* Δ*aceE* Δ*fdhF* background		
	15 h	24 h
fdhA	1.30	0.98
hnr	1.23	0.86
Δ*gdhA* background		
	15 h	24 h
clpXP	1.37	1.59
hnr	1.32	1.23
lipB	1.30	1.27
hycI	1.24	1.01
ygjP	1.22	1.23
fdhA	1.21	1.28
yagR	1.20	1.29
gntK	1.17	1.19
pflB	1.03	1.32
glnE	1.01	1.35
ackA	0.91	1.18
modA	0.90	1.43
Δ*yjfP* background		
	15 h	24 h
clpXP	1.46	1.60
fdhD	1.28	1.12
cyaA	1.22	1.09
nuoK	1.16	1.08
aspC	1.16	1.03
glnE	1.15	1.06
fdhA	1.08	1.16
feoB	0.94	1.20
clp	0.90	1.24

Table 6.9: Fold improvement in lycopene production by identified gene knockouts.

Δ*hnr* background		
	15 h	24 h
yliE	2.55	1.38
sohA	2.11	1.26
cyaA	2.04	1.22
pitA	1.87	1.38
fdhA	1.80	1.12
yjhH	1.70	1.12
yfcC	1.67	1.26
aspC	1.59	1.03
yibD	1.52	1.23
lysU	1.48	1.18
yedN	1.44	1.12
yebB	1.42	1.09
fumA	1.40	1.10
csdA	1.40	1.01
ycfZ	1.38	1.26
crcB	1.38	1.14
yaiD	1.33	0.90
ydeN	1.31	1.25
Δ_{P}*yjiD* background		
	15 h	24 h
ybaS	0.73	1.28
appY prom	1.16	1.03
clpP	1.53	1.33
glxR	0.69	1.30
Δ*gdhA* Δ*aceE* Δ_{P}*yjiD* background		
	15 h	24 h
nadA	0.96	1.21
evgS	1.07	0.93
stpA	1.13	0.94
ackA	0.79	0.93
moeA	1.13	0.94
pflB	1.03	0.79
pstC	1.30	1.33

Table 6.9: Fold improvement in lycopene production by identified gene knockouts.

6.2 Creating a search network diagram

The various gene knockout combinations yielding increased lycopene yield may be represented on a gene knockout search network. In such a representation, the circles, or nodes, represent gene knockouts and the edges that connect them are directional and represent trajectories (or possible paths) for increasing lycopene production. As an example, the metabolic landscape of **Figure 5.5** may be represented in the search network depicted in **Figure 6.1**. Embedded in this representation are genotypes for the best strains throughout the metabolic landscape. In this type of figure, the circles, or nodes, represent gene knockout targets. The arrows connecting these nodes are directional and represent trajectories for increased lycopene production. For example, following the various trajectories can lead, starting from the parental strain node (labeled "none"), in three steps to either of the two global triple knockout strains maxima strains, Δ*gdhA* Δ*aceE* Δ*fdhF* and Δ*gdhA* Δ*aceE* Δ$_P$*yjiD*. Furthermore, it is possible to find the three combinatorial gene knockout targets, Δ*hnr*, Δ*yjfP*, and Δ$_P$*yjiD* which are highlighted in **Table 6.1**. This search network can be expanded using the eight combinatorial gene knockout searches described in **Tables 6.1 – 6.8**. The complete gene knockout search network is presented in **Figure 6.2**. The topology of this network is discussed in more depth in the following sections to extract information regarding the importance of search trajectory on the exploration of a metabolic landscape.

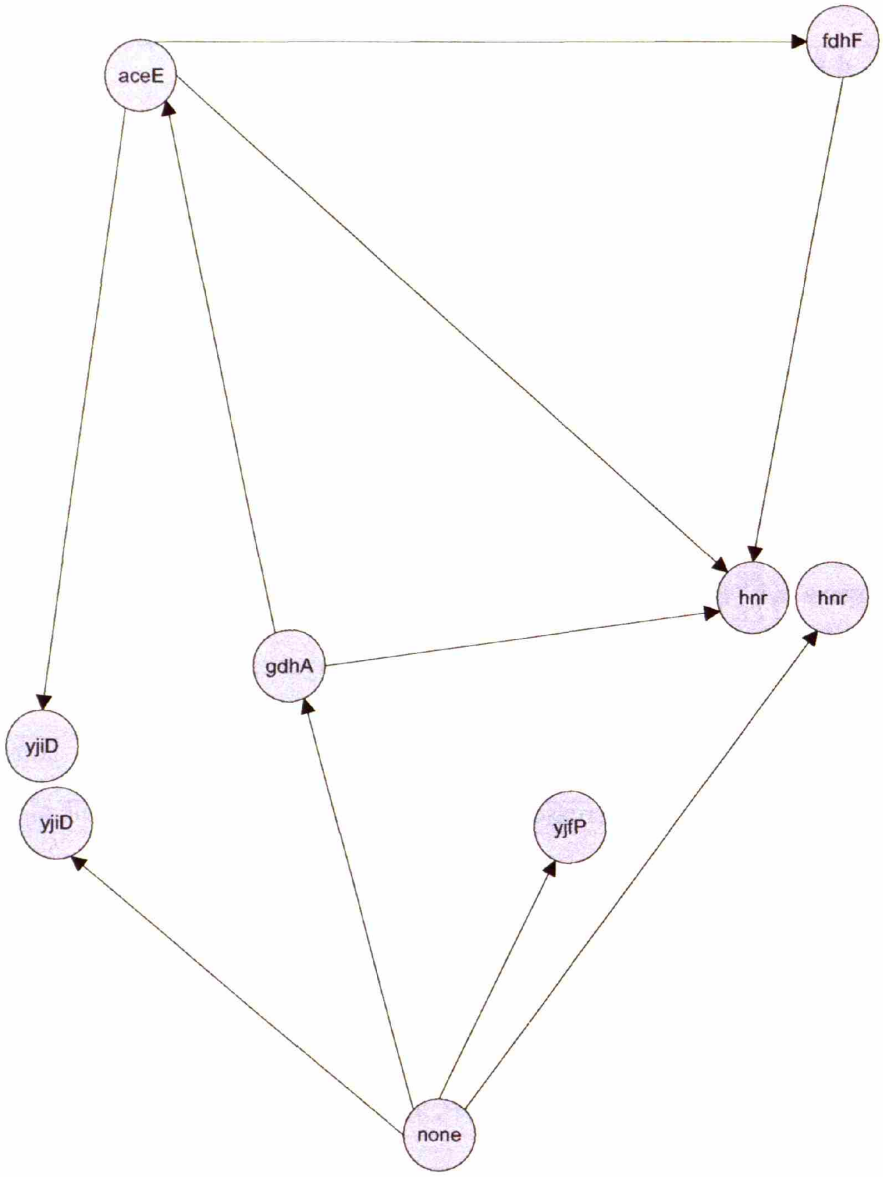

Figure 6.1: Original search network diagram for the metabolic landscape. The gene knockout search network for the metabolic landscape in **Figure 5.5** is presented. Following the various trajectories can lead to the two global maxima strains, Δ*gdhA* Δ*aceE* Δ*fdhF* and Δ*gdhA* Δ*aceE* Δ$_{P}$*yjiD*. Furthermore, it is possible to find the three combinatorial gene knockout targets, Δ*hnr*, Δ*yjfP*, and Δ$_{P}$*yjiD*.

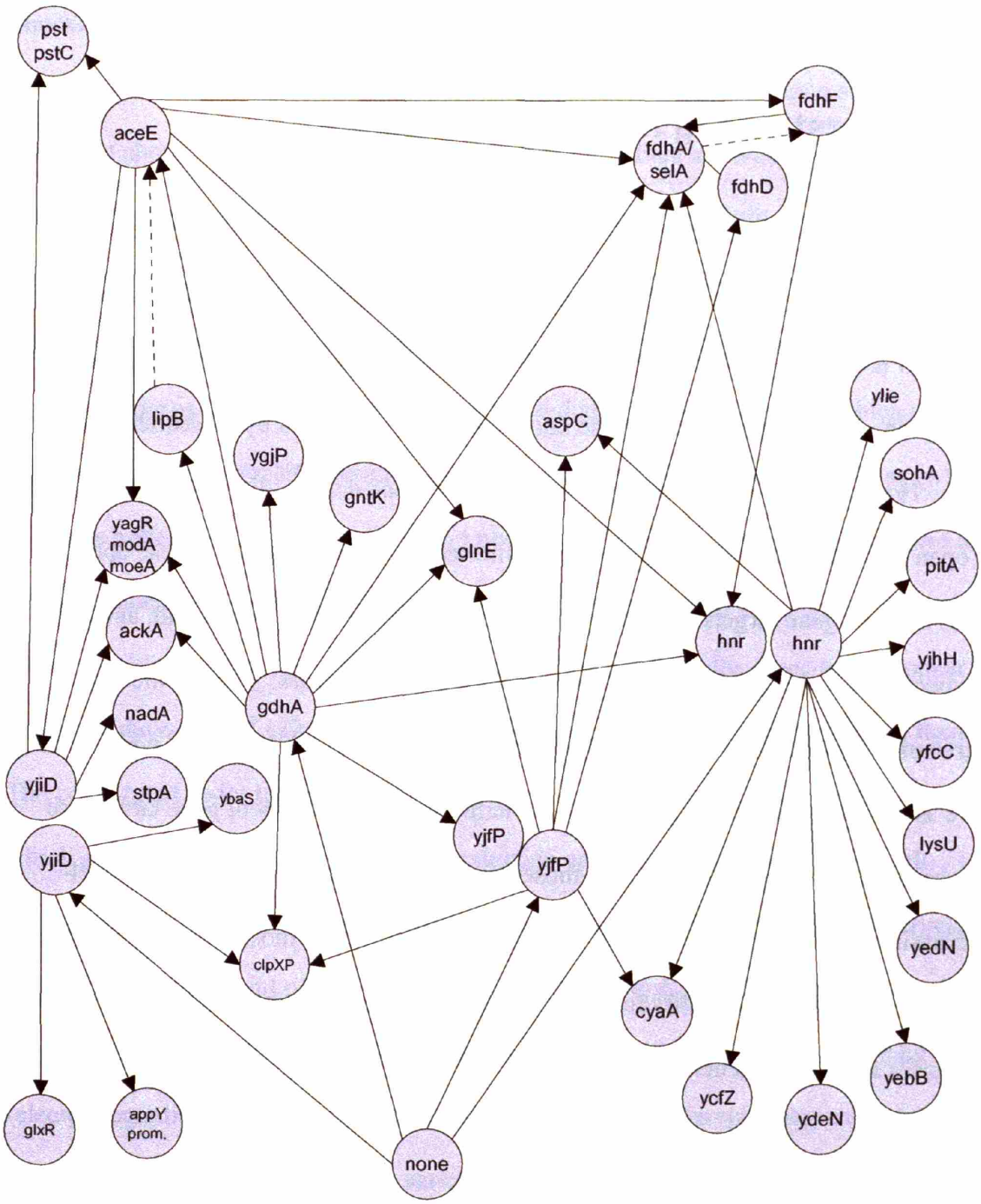

Figure 6.2: Complete search network diagram for the metabolic landscape.

Inclusion of all the targets from **Tables 6.1 – 6.8** lead to a complete gene knockout search network provided here. Dashed lines represent two literature annotated protein-protein interactions.

6.3 Understanding network complexity

Immediate investigation into the topology of the search network (**Figure 6.2**) presents several observations of importance to metabolic engineering. The landscape presented is quite complex with many possible paths leading to increased lycopene production. In total, there are 37 gene knockout nodes connected by 52 search trajectory edges. However, all nodes are not connected to the same degree. For example, there are 17 open set nodes (nodes which are only connected to one other node). These open set nodes indicate gene targets that only arise in a particular genetic background, and point to the specificity and importance of genotype on the search network. Despite the high degree of complexity apparent from this graph, there are several "gateway" nodes which serve as critical points in the network. Furthermore, a wide variety of gene targets are represented in this network including those that are model-accessible, mostly model-inaccessible, and model inaccessible. These classifications are mostly created based on available annotation and ability to create adequate models. Members from each of these categories are described below.

6.3.1 Gateway nodes

Despite the complexity seen in the search network diagram of **Figure 6.2**, most of this complexity can be accessed by choosing trajectories which pass through a few "gateway" nodes. These nodes are important gene knockout targets that exhibit unique connectivity properties within the network. Specifically, these nodes represent gene

knockout targets that are either universal (independent of genetic background) or ones that allow for further exploration of complexity (allows for access to open set nodes). As discussed above, there are a number of open set nodes which are only connected to one particular gene knockout genotype. These nodes represent knockouts which are specific to the genotype under study. Conversely, there are a number of highly connected nodes which can impart a divergent, convergent, or both structure within the network topology and thus serve as important gateway nodes. Included in this set are *gdhA* (which serves as a divergent node) the *fdh* operon (which serves as a convergent node), and *hnr* (which serves as both a convergent and divergent node). The divergent nodes (such as *gdhA* and *hnr*) are gene knockout targets which allow for the exploration of novel genetic targets only available once the first perturbation has been made. Visually, these nodes appear similar to the center of a tire with many spokes (new trajectories) eminating from them. The convergent nodes (such as *fdh* operon and *hnr*) illustrate gene knockout targets which are more universal for lycopene production and thus were identified through several searches in different genetic backgrounds. These highly-connected convergent nodes illustrate the point of recurrence in the search network. In particular, knockout targets which show a high amount of recurrence are universal targets which will be selected regardless of the genetic background. Divergent nodes represent essential gene knockouts which enable access to a variety of diverse gene knockout targets. In particular, these knockout targets are key nodes for the construction of improved strains. Collectively, these central, or gateway nodes present key gene knockout targets universally required for lycopene production in *E. coli*. More importantly, these small number of gateway nodes illustrate that despite the high degree of complexity and non-

linearity, there are still a small subset of genes which can provide access to all points in the search network.

6.3.2 Model accessible nodes

Several important metabolic pathways, which may be modeled, are overrepresented in this search network. Furthermore, these overrepresented metabolic pathways correspond with the the gene knockout targets identified by the stoichiometric (systematic) analysis previously conducted. Therefore, the results of this transposon mutagenesis study validate the method of global stoichiometric modeling to identify critical regions of metabolism related to a specific phenotype (in this case, lycopene production). In particular, glutamate metabolism appears in many genotypes through the knockout of genes such as *gdhA*, *ybaS* (Reed et al., 2003), *glnE*, and *aspC*. Pyruvate metabolism appears in many genotypes through the knockout of genes such as *aceE* and *lipB* (Jordan & Cronan, 2003). The formate dehydrogenase complex arises in most backgrounds through knockouts in various genes of the *fdh* operon. Furthermore, genes encoding selenocysteine biosynthesis such as *fdhA* were prevalent in this search and are directly related to the activity of *fdhF*, a protein that requires selenocysteine incorporation for function (Leinfelder et al., 1988; Zinoni et al., 1986). Finally, several phosphate transport gene knockouts appeared through the search with the appearance of genes such as *pstC*, *pst*, and *pitA*. These genes could be embodied in models since their function are of stoichiometric nature.

It is important to note that in the background of a *gdhA* knockout, genes related with *aceE* were discovered. In the background of a *gdhA* and *aceE* double knockout,

genes related to *fdhF* were discovered. These results echo the systematic search for stoichiometric targets using the models discussed in Chapter 4. These targets indicate that key strategies to increasing carotenoid production would involve these three sets of metabolic targets.

6.3.3 Mostly model inaccessible nodes

Several important regulators were found to be overrepresented in this search network, many of which overlap with the metabolic targets described above. These targets may be of known function, however they are often regulators with pleiotropic effects which cannot be adequately embodied in a cellular model. In particular, glutamate metabolism regulators were found such as *glnE*. This target highlights the importance of the glutamate node in the search network. Global genetic regulation was altered by the identification of genes such as *hnr* and *clpXP* which can presumably act through the increased steady state levels of σ^S (*rpoS*). The importance of *rpoS* for carotenoid production in *E.* coli has been previously reported (Becker-Hapak et al., 1997). Starvation response was also affected through the identification of a promoter knockout of the *appY* gene. Furthermore, many putative regulators such as *ygjP*, the promoter region of *ycfZ*, *stpA* were found to be altered. These nodes are highly pleiotropic as genetic regulators and could lead to the identification of many genes related with lycopene production.

6.3.4 Model inaccessible nodes

Finally, a total of 38% of the gene knockout nodes present in the search network were of unknown or putatively assigned function. This indicates the potential that there are several unique modes of action uncovered through this search. Furthermore, these gene knockout targets often provided the most significant increases in lycopene production. The inability to obtain annotations for these genes, and their subsequent importance on lycopene production, highlights the difficulty of being able to use models for the improvement of strains.

6.4 Further characterization of strains

Several strains of interest from this search network were investigated and characterized further to determine their lycopene production potential. In particular, the gene knockout targets of *yliE*, *pstC*, *pitA, pst operon*, and *clpXP* were investigated in specific genetic backgrounds. The gene *yliE* is a hypothetical protein which was found in the *hnr* background and provided a very significant increase in lycopene production. The phosphate transporter genes (*pstC*, *pitA,* and *pst operon*) were identified in three separate genotypes, Δ*gdhA* Δ*aceE*, Δ*gdhA* Δ*aceE* Δ$_{P}$*yjiD*, and Δ*hnr*. Finally, the ATP-dependent proteases, *clpXP*, were found in three separate genotypes, Δ*gdhA*, Δ$_{P}$*yjiD*, and Δ*yjfP*, and serve as an additional example of recurrence of targets in this search network.

Figure 6.3 presents the lycopene levels at 15 and 24 hour timepoints in 1xM9 medium compared to the previously identified global maxima strains. Comparing these

strains to the previous global maxima illustrates that different genotypes are able to yield the same phenotype of lycopene production. Despite many similarities, the premier strain from this analysis was the Δ*hnr* Δ*yliE* construct which outperformed all previous strains (including the previously identified global maxima) at the 15 and 24 hour timepoints.

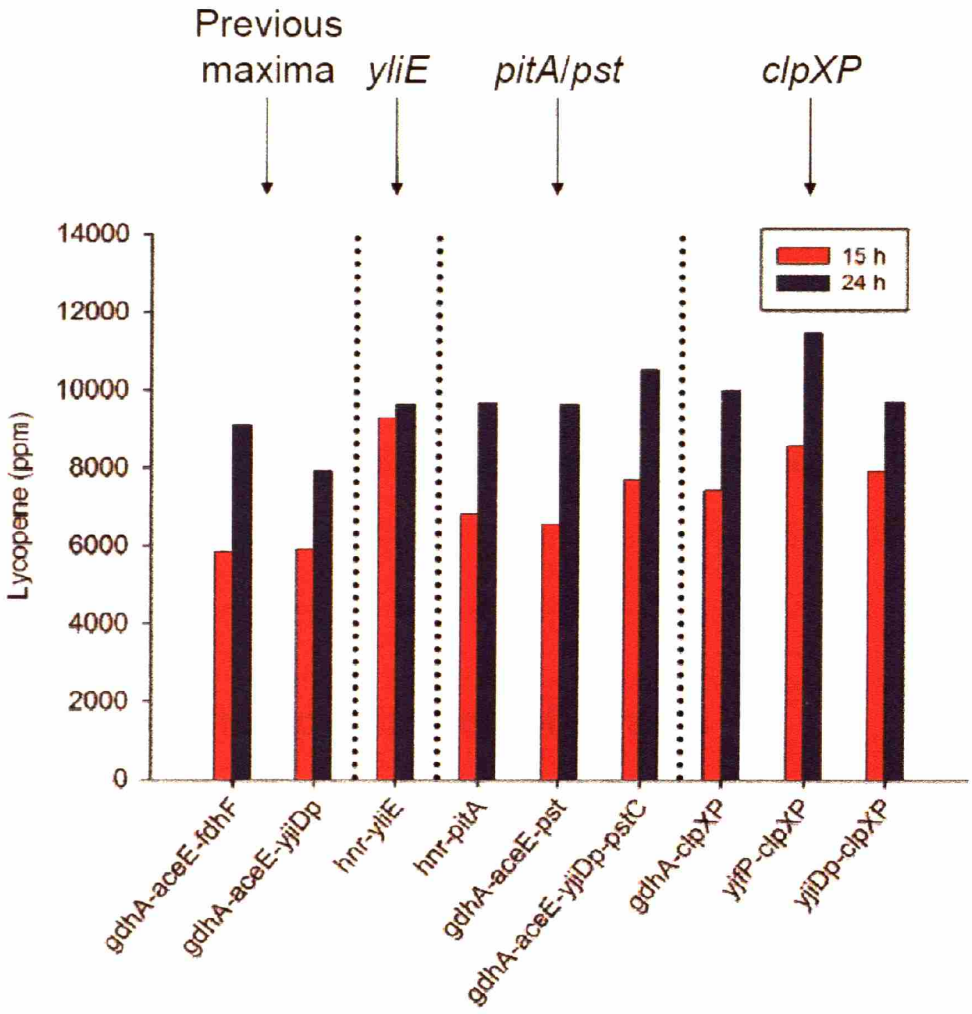

Figure 6.3: Lycopene production of selected strains at 15 and 24 hours in 1xM9.

Lycopene production levels of selected strains from the search network are presented at 15 and 24 hours when cultured in a 1x M9 minimal medium.

6.4.1 *yliE* investigation

The Δ*hnr* Δ*yliE* strain appeared to be superior to the previously identified global maxima strains as shown in **Figure 6.3**. This strain was then analyzed in 2x M9 medium, which is known to enhance lycopene production (Alper, Miyaoku, & Stephanopoulos, 2005). Under these conditions **(Figure 6.4)**, the Δ*hnr* Δ*yliE* strain showed a slightly increased lycopene production at the 15 hour timepoint in comparison to the previously identified global maxima. However, this strain was suboptimal when comparing production levels at the 24 hour timepoint. It is important to note that the Δ*hnr* Δ*yliE* strain was identified based on its performance in 1x M9 medium. These results indicate that the gene knockout targets are specific to the culturing environment in which they are selected. Often, such a relationship is termed a "Gene x Environment" effect.

Finally, the Δ*yliE* genotype was investigated in various other genotype backgrounds to assess broad applicability. **Figure 6.5** presents the results of a *yliE* knockout in various genetic backgrounds. These results indicate (and reconfirm) *yliE* as an open set node which is a genetic target specific to the genotype of the strain. In particular, Δ*yliE* was not able to positively influence lycopene production in a significant way in any background except for Δ*hnr*. These results point to the specificity of identified gene targets for the genotype and is often termed a "Gene x Genotype" effect. This relationship greatly confounds searches for gene targets and prevents the ability to transfer identified targets to any genotype of choice.

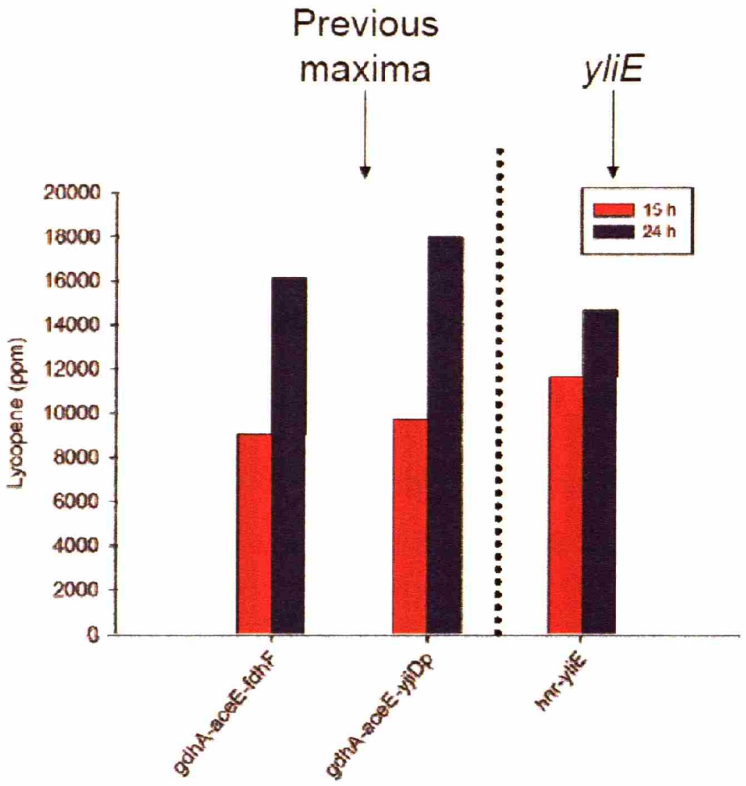

Figure 6.4: Comparison of Δ*hnr* Δ*yliE* to previous maxima at 15 and 24 hours in 2xM9. While the Δ*hnr* Δ*yliE* strain outperformed the maxima in 1x M9 medium, these results indicate the importance of environmental/culturing conditions on the identification of gene targets.

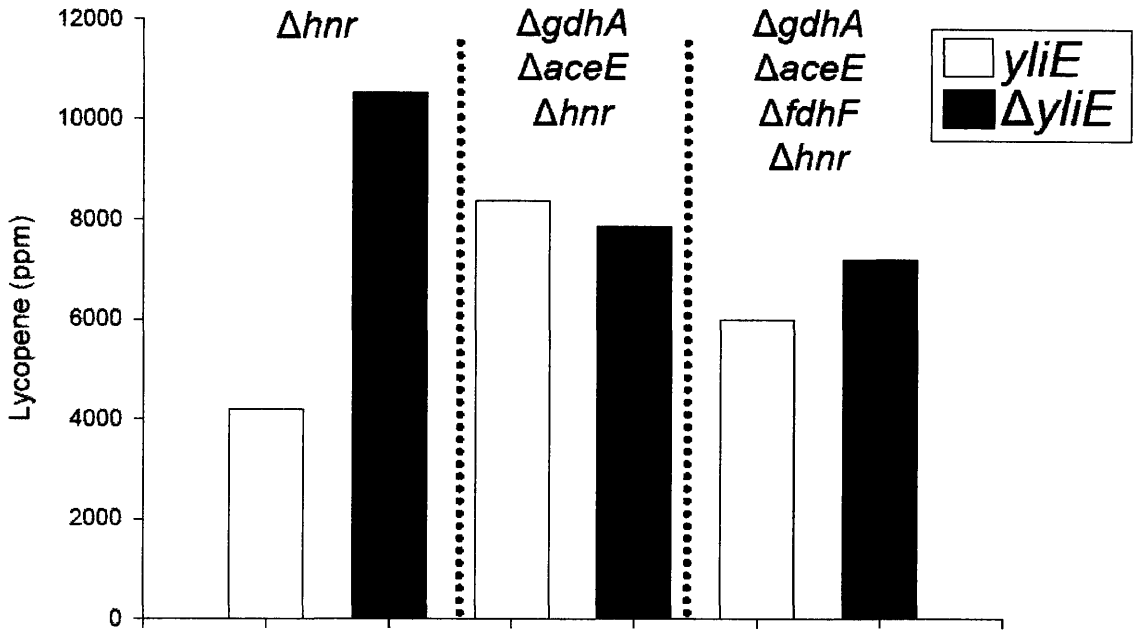

Figure 6.5: Impact of Δ*yliE* in other genotypes. The Δ*yliE* genotype was transferred to other *hnr* knockout strains. However, the impact of this gene knockout was only pronounced in the background for which it was identified, the Δ*hnr* strains. These results illustrate the specificity of some targets for the genotype and reaffirm *yliE* as an open-set node in this search network.

6.5 Summary

Cumulatively, these results point to the existence of multiple paths which can traverse the metabolic landscape. This fact leads to the finding that many diverse genotypes can yield the same overall phenotype. Recurrence of gene targets and modes of action such as those with glutamate metabolism and the *fdh* operon point to significant portions of metabolism responsible for the carotenoid production phenotype. However, a significant number of gene targets (38%) were uncharacterized and unannotated which makes mechanistic understanding difficult.

In general, this investigation illustrated a high amount of complexity seen in metabolic landscapes. However, while the gene knockout search network was non-linear and complex to traverse, most of the interesting trajectories were controlled through a small subset of "gateway" nodes such as those passing through glutamate and the FDH operon. The high amount of unknown nodes (38%) and genes of regulatory nature highlight the limitations of current model-based approaches which can only access a small portion of this metabolic landscape. Finally, it is evident from this analysis that there exists a many-to-one mapping of genotype to phenotype, which can complicate the process of searching and optimizing metabolic landscapes.

The results presented in this chapter and the previous two indicate that complexity increases with number of components samples. Moving from the definition of single and multiple gene knockouts afforded by a stoichiometric model to the set of genes spanning the metabolic landscape of systematic and combinatorial targets to the creation of the search network required increasing effort for target identification. With these additional

components came additional improvements in phenotype, but at the cost of increased complexity and non-linearity. During the identification of systematic targets, a method of steepest ascent was taken to identify multiple targets. Through an exhaustive search, it was shown that this method was applicable. However, the superposition of regulatory and unknown factors in these strains makes taking such a trajectory impossible. These conclusions are highlighted by comparing phenotype optimization of the exhaustive search to that of following the "greedy algorithm". Using a greedy algorithm for optimization would have failed in optimizing this metabolic system, leading to only the 15th best construct (~25% lower than highest strain). While these results only dealt with gene knockout targets, the same analysis, and results, would be seen for any perturbation which acts at the single-gene level.

As a result of these analyses, it becomes evident that metabolic landscapes are too complex and nonlinear to *efficiently* and *completely* probe through *single gene modifications* linked with a search strategy. As a result, it is necessary to create a number of tools for engineering cells at the *global* level. Necessary components to this end include the ability to engineer and tune genetic control of single genes aided through the development of a functional promoter library. In addition, tools are required which can afford multiple modifications to genes simultaneously. The development and implementation of these tools are discussed in the following chapters and can lead to further optimization of these metabolic landscapes.

Chapter 7

Promoter Engineering

7.1 Motivation

Effective probing of a metabolic landscape requires not only gene deletions, but also varying (or tuning) the expression level of a gene of interest. In most previous studies, gene function is typically evaluated by sampling the continuum of gene expression at only a few discrete points corresponding to gene knockout or overexpression, often decided by the availability of varying expression plasmids. However, such a characterization is incomplete and inadequate for creating all the possible gene expression levels necessary in a global manner. To address this issue, the tool of promoter engineering was developed. This chapter discusses the creation and implementation of a library of engineered promoters of varying strength obtained through mutagenesis of a constitutive promoter. A multi-faceted characterization of the library, especially at the single-cell level to ensure homogeneity, permitted quantitative assessment correlating the effect of gene expression levels to improved growth and product formation phenotypes in *E. coli*. Integration of these promoters into the chromosome can allow for a quantitative, accurate assessment of genetic control. To this

end, we utilized the characterized library of promoters to assess three phenotypes: (1) the impact of phosphoenolpyruvate carboxylase (PPC) levels on growth yield, (2) deoxy-xylulose-P synthase (DXS) levels on lycopene production, and (3) the impact of *ppc* knockdown on lycopene yield. Collectively, these examples illustrate that optimal gene expression levels are variable and dependent on the genetic background of the strain. As a result, tools such as promoter engineering which allow for a wide range of expression levels constitutes an integral platform for functional genomics, synthetic biology, and metabolic engineering endeavors.

7.2 Background

Protein engineering via directed evolution and gene shuffling (Glieder, Farinas, & Arnold, 2002; Stemmer, 1994) has been extensively applied for the systematic improvement of protein properties such as antibody binding affinity (Boder, Midelfort, & Wittrup, 2000), enzyme regulation (Nelms et al., 1992), and increased or diverse substrate specificity (Fa et al., 2004). A similar approach whereby continuously improved mutants are generated along a selection-defined trajectory in the sequence space can also be applied for the systematic improvement or modification of other types of biological sequences, e.g. ribozymes (Ferguson et al., 2004; Tao, Jackson, & Cheng, 2005). This work illustrates that promoters can also be engineered via directed evolution to achieve precise strengths and regulation, and, by extension, can constitute libraries exhibiting broad ranges of genetic control.

Typically, the deletion (Zhou et al., 2003) and the strong over-expression (Nishino, Inazumi, & Yamaguchi, 2003) of genes have been the principal strategies for elucidation of gene function. These two methods sample the continuum of gene expression at only a few discrete points, determined by experimental feasibility (Jana & Deb, 2005) and not necessarily biological significance. Thus, the full dependency of phenotype on gene expression may not be accessible due to the limitations inherent in these methods. Gene expression is controlled by a number of factors in the cell including promoter strength, cis- and trans-acting factors, cell growth stage, the expression level of various RNA polymerase-associated factors, and other gene-level regulation. Of course, gene expression may not always correspond with enzymatic activity given protein level regulation which may also be present. Nevertheless, several groups have attempted to control gene expression through the creation of promoter libraries (Jensen & Hammer, 1998; Jorgensen et al., 2004; Khlebnikov et al., 2000). This chapter discusses a *fully-characterized*, homogeneous, broad-range, functional promoter library and *demonstrates* its applicability to the analysis of such a genetic control. By characterizing the strength of these promoters in a quantitative manner with various metrics and subsequently integrate these constructs into the genome, it is possible to deduce the precise impact of the gene dosage on the desired phenotype.

An alternative method for controlling gene expression is through the use of a single inducible promoter tested at various levels of inducer. While inducible promoters allow for a continuous control of expression at the macroscopic level, practical applications of these systems are limited by prohibitive inducer costs, hypersensitivity to inducer concentration, and transcriptional heterogeneity at the single-cell level

(Mnaimneh et al., 2004; D. A. Siegele & Hu, 1997). The latter factor in particular, can limit the effect of inducers in a culture to a simple increase of the number of cells expressing the gene of interest instead of the overexpression of the gene in all cells. Inducible systems are suitable in certain applications (e.g. recombinant protein overproduction) (San et al., 1994); however, the elucidation of gene function and genetic control on phenotype requires well characterized promoter libraries which behave in a similar manner at the single cell level. As a result, the creation of a promoter library based on a constitutive promoter would eliminate the need to regulate inducer concentrations and avoid heterogeneities in cellular response.

7.3 Implementation

A derivative of the constitutive bacteriophage P_L- λ promoter (Lutz & Bujard, 1997) was mutated through error-prone PCR (Zaccolo et al., 1996), cloned into a reporter plasmid upstream of a low-stability GFP gene (Andersen et al., 1998), and screened in *E. coli* based on the fluorescence signal in a glucose minimal medium, supplemented with 0.1% casamino acids to attenuate GFP toxicity. Nearly 200 promoter mutants, spanning a wide range of GFP fluorescence, were selected. Many of these initially screened promoters exhibited large variations in fluorescence between several trials or did not have an acceptable *single-cell* level homogeneity. Twenty-two mutants were finally chosen to form a functional promoter library based on reproducible and homogeneous single-cell fluorescence distributions as measured by flow cytometry. **Figure 7.1** illustrates the process of creating and subsequently selecting these promoters. The functional promoter

library was analyzed using flow cytometry. The relative average geometric mean fluorescence of the members of the library are illustrated in **Figure 7.2** and are seen to exhibit a nearly 3 log fold range after 14 hours in a minimal media with 0.1% Casamino acids.

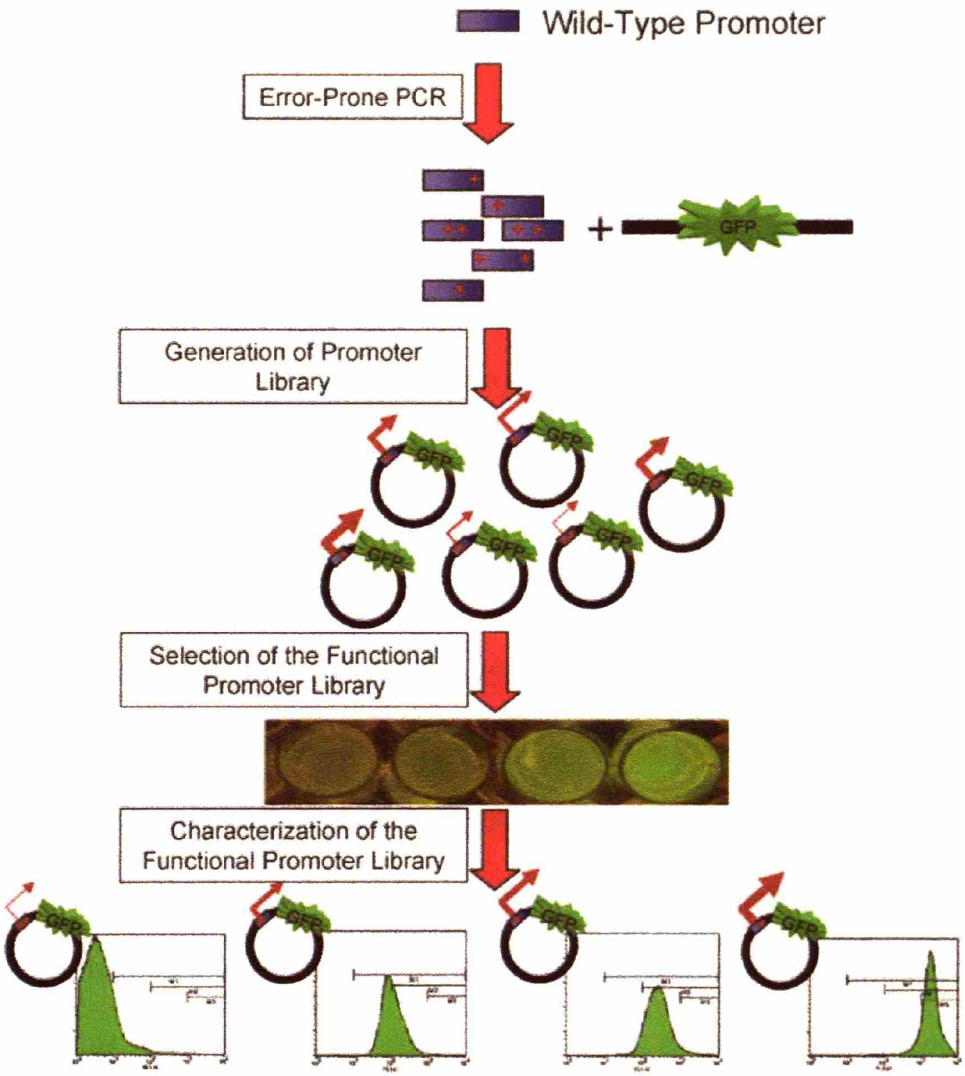

Figure 7.1: Generation of the functional promoter library. A variant of the constitutive bacteriophage P_L- λ promoter was mutated through error-prone PCR, used in a plasmid construct to drive the expression of *gfp*, then screened based on fluorescence of colonies. The chosen constructs have a wide range of fluorescence both on a culture-wide level and on a single-cell level as illustrated by representative flow cytometry histograms at the bottom. All of the selected promoters have a uniform expression level on a single cell level as measured by GFP signal.

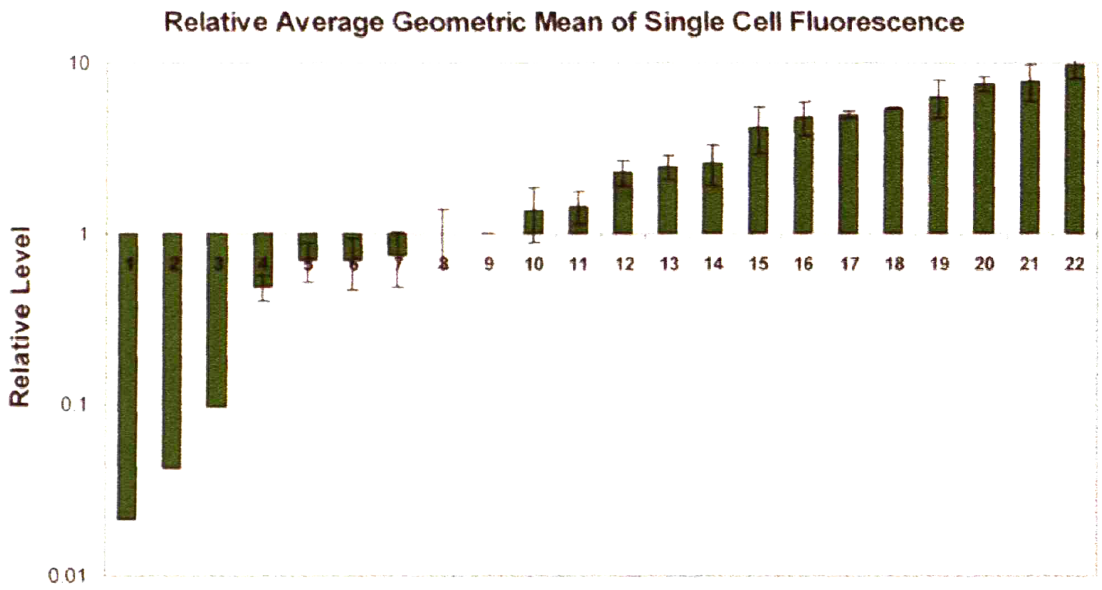

Figure 7.2: Flow cytometry analysis of the functional promoter library. Members of the functional promoter library are assayed for GFP fluorescence using flow cyteomery and are shown to exhibit a nearly 3 log fold range after 14 hours in a minimal media with 0.1% Casamino acids.

7.3.1 Multi-faceted characterization

In light of the uncertainty surrounding the concept of promoter strength (Horn & Wells, 1981) and the poor reliability of single reporter-gene-based systems, we performed a multi-faceted characterization of each library member. We first determined the promoter strength in the library strains (in units of GFP fluorescence per cell per hour) by measuring culture fluorescence and using a dynamic equation balancing GFP production and degradation (Leveau & Lindow, 2001). In particular, fluorescence readings taken during the exponential growth phase were plotted as a function of turbidity. The best-fit slope to this line represents the exponential-phase steady-state concentration of GFP, f_{SS}. Because f_{SS} is affected by the cell growth rate, oxygen-dependent maturation constant of GFP, and the protease-mediated degradation of GFP as well as the promoter-driven synthesis of new GFP, it is not a suitable metric for promoter strength. Instead, we used a previously published dynamic model that accounts for all of these factors. Under this model, shown in **Figure 7.3**, and under the assumption that the rate constant of protease-mediated degradation is the same for mature GFP as its precursor polypeptide, P, the rate of promoter-driven production of GFP may be calculated. Through replicate culturing, the promoter strength of the library members was found to span a 196-fold range with a mean spacing of 29% between adjacent members.

Next, to characterize the promoter library directly at the transcriptional level, we measured the relative mRNA levels of *gfp* transcripts in the above cultures by quantitative RT-PCR. The high correlation between fluorescence and mRNA level

confirmed that expression was transcriptionally controlled. The mRNA level spanned a 325-fold range with a mean spacing of 32% between adjacent members. We then formed an "average promoter strength metric" for each promoter by averaging the scaled mRNA and fluorescence data.

Finally, to verify the constitutive nature of all the promoters, each was redeployed into a new construct driving the reporter gene *cat*. Cultures bearing these constructs were assayed for resistance to chloramphenicol on a rich, solid-phase medium. The MIC spanned a 26-fold range with a mean spacing between MIC values of 17% (which is biased due to a discrete levels of chloramphenicol tested). The results of these characterizations are shown in **Figure 7.3**.

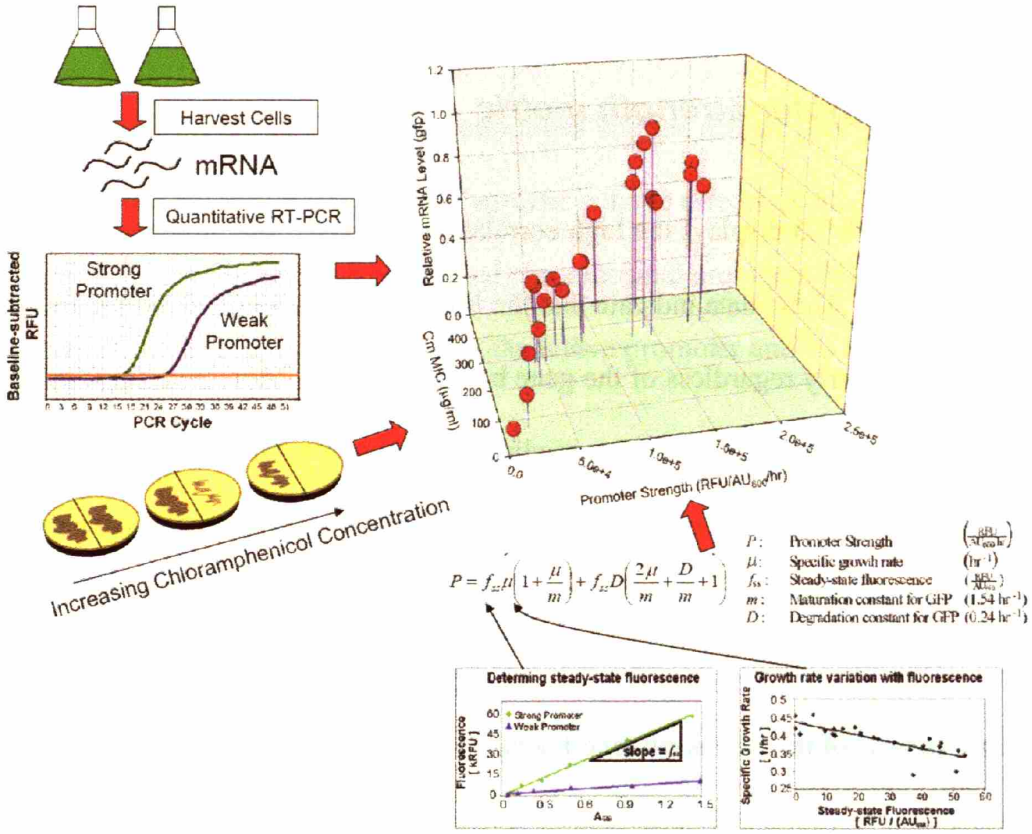

Figure 7.3: Comprehensive characterization of the promoter library. Several orthogonal metrics were employed to characterize the promoter library and quantifying the transcriptional activity of the promoters: (1) The dynamics of GFP production based on fluorescence, (2) measurement of the relative mRNA transcript levels in the cultures, and (3) testing of the MIC for chloramphenicol in an additional library of constructs where the promoter drove the expression of CAT. The overall strong correlation between the various metrics suggests a broad-range utility of the promoter library for a variety of genes and conditions.

7.3.2 Promoter strength metric

Figure 7.3 displays the high correlation between these three metrics of promoter performance. These data indicate that the library exhibits a high dynamic range which behaves similarly regardless of the gene being regulated. Moreover, these conditions test the promoter library in contrasting medium and growth environments (liquid minimal medium vs. solid complex medium) further underscoring the constitutive nature of the library promoters. The disparity between the number of initially and finally selected promoters illustrates the need for a comprehensive analysis of the promoters. While many subsets of mutations can elicit a change in promoter strength, not all are guaranteed to lead to a reproducible, homogenous, and linear relationship between promoter strength and reporter. Relying solely on bulk culture-based measurements can lead to misclassification of the behavior of the promoter at the single cell level and thus complicate quantitative gene expression studies, such as those performed in this study. As a result, a promoter strength metric was created which served as an average of relative strength of these promoters as judged by these three assays. It is necessary to use multiple assays since various measurements using reporter genes (such as GFP or CAT) have biases which may confound future analysis of gene function.

7.4 Applications

The functional promoter library was introduced into the cell for precise transcriptional control for the investigation of specific genetic effects on a cellular phenotype in *E. coli*. To this end, we performed chromosomal promoter delivery into the region upstream of the targeted gene, replacing the native promoter and its inherent regulation modality.

7.4.1 Growth yield and ppc activity

Enabled with a fully characterized library, it was possible to asses how the expression of *ppc* impacted biomass yield from glucose. This gene expresses phosphoenol pyruvate (PEP) carboxylase, a key anaplerotic enzyme. A *ppc* knockout is lethal for *E. coli* in glucose minimal medium (McAlister, Evans, & Smith, 1981). Furthermore, overexpression of this gene has been shown to improve the growth yield on glucose (Liao, Chao, & Patnaik, 1994). These data imply two possibilities: either biomass yield is a monotonically increasing function of *ppc* expression or there exists a particular *ppc* expression level which maximizes yield. To address this issue, *E. coli*'s native *ppc* promoter was replaced with varying-strength promoter-*ppc* constructs, and these mutants were cultured while biomass and glucose concentrations were periodically monitored. **Figure 7.4** presents the exponential-phase biomass yields as a function of the average promoter strength metric. Increasing *ppc* levels have a positive effect on the biomass yield only to a certain point. This increase reaches a plateau, and further increases in the *ppc* level have a *negative* effect on the biomass yield. These results

illustrate an optimum in the expression level of *ppc* that is *above* that found from endogenous expression. Possible reasons why ever-increasing *ppc* levels lead eventually to a decrease in yield include the metabolic burden of severe overexpression of *ppc* or, more likely, the creation of a futile ATP-wasting cycle in metabolism where PEP is converted to oxaloacetate by *ppc* and back again by *pck*, the gene for PEP carboxykinase.

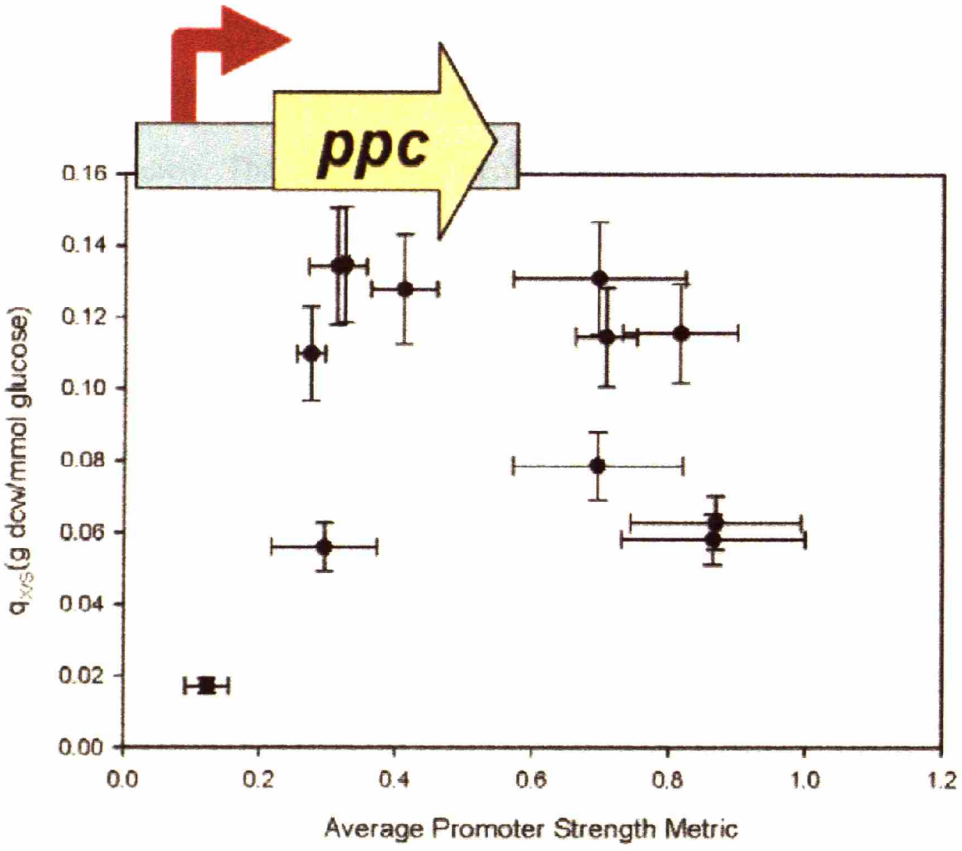

Figure 7.4: Growth yield and *ppc* activity. Selected promoters were integrated into the promoter region of *ppc* and strains were cultured in M9-minimal media with only glucose as the carbon source. While the knockout of *ppc* is lethal in glucose media, there is a clear maximum yield from glucose and thus an optimal expression level of *ppc*.

7.4.2 Lycopene yield and dxs activity

Kinetic control of metabolic pathways is often distributed and dependent on the expression level of several genes within the pathway (Stephanopoulos & Vallino, 1991). The gene *dxs* represents the first committed step in isoprenoid synthesis in *E. coli* and has been implicated in control of lycopene production (Seon-Won Kim & Keasling, 2001); however, the quantitative nature of this control was unclear, and promoter delivery experiments also allowed for the quantification of this control in multiple backgrounds. Here, volumetric productivity of lycopene accumulation in glucose medium was investigated as a function of the expression levels of the *dxs* gene in two different *E. coli* strains: the wild-type K12 strain and a previously engineered strain which already produces lycopene in high titers (Alper et al., 2005b). **Figure 7.5** shows the lycopene production in these *dxs* constructs in a wild-type (K12) background. Elevating *dxs* expression increases lycopene accumulation only until a certain point. Beyond this optimum, increased *dxs* expression is detrimental for lycopene production. Finally, the strength of the native *dxs* promoter can be inferred from this analysis as is illustrated on the graph.

In contrast to the above results, a linear relationship was obtained when similar promoter-*dxs* constructs were placed in an engineered strain (Alper et al., 2005b) overexpressing downstream genes in the isoprenoid pathway (*ispFD* and *idi*). **Figure 7.6** illustrates a nearly linear response of lycopene production to varying levels of *dxs* expression, suggesting that in the new genetic background, *dxs* has become rate-limiting.

In the case of wild-type *E. coli*, an optimal *dxs* expression was again apparent, similar to the analysis with *ppc*. Past the optimum, increasing *dxs* expression lowers

lycopene yield, presumably due to the inadequate activity of downstream enzymes in the isoprenoid pathway and resulting toxic buildup of DXP. In contrast, in a strain already engineered to overexpress *idi*, *ispF*, and *ispD*, downstream genes in lycopene biosynthesis, no maximum is apparent. A linear response to an enzyme concentration is expected for rate-controlling genes exhibiting a high flux control coefficient for a given pathway (Kacser & Acerenza, 1993), suggesting that even at the highest expression levels examined in this study, the *dxs*-catalyzed reaction is rate-limiting for lycopene biosynthesis. We also note that cell density in both strains was greatly reduced in the constructs harboring low-strength promoters, which was expected, as *dxs* is an essential gene. A significant step in performing these quantitative functional genomics studies is creating a reliable, characterized promoter library for which confidence in the cellular gene expression level may be placed. When this initial step is established, it is possible to quantitatively analyze the control a single enzyme exerts in a given pathway of interest, exemplified by the *dxs* example.

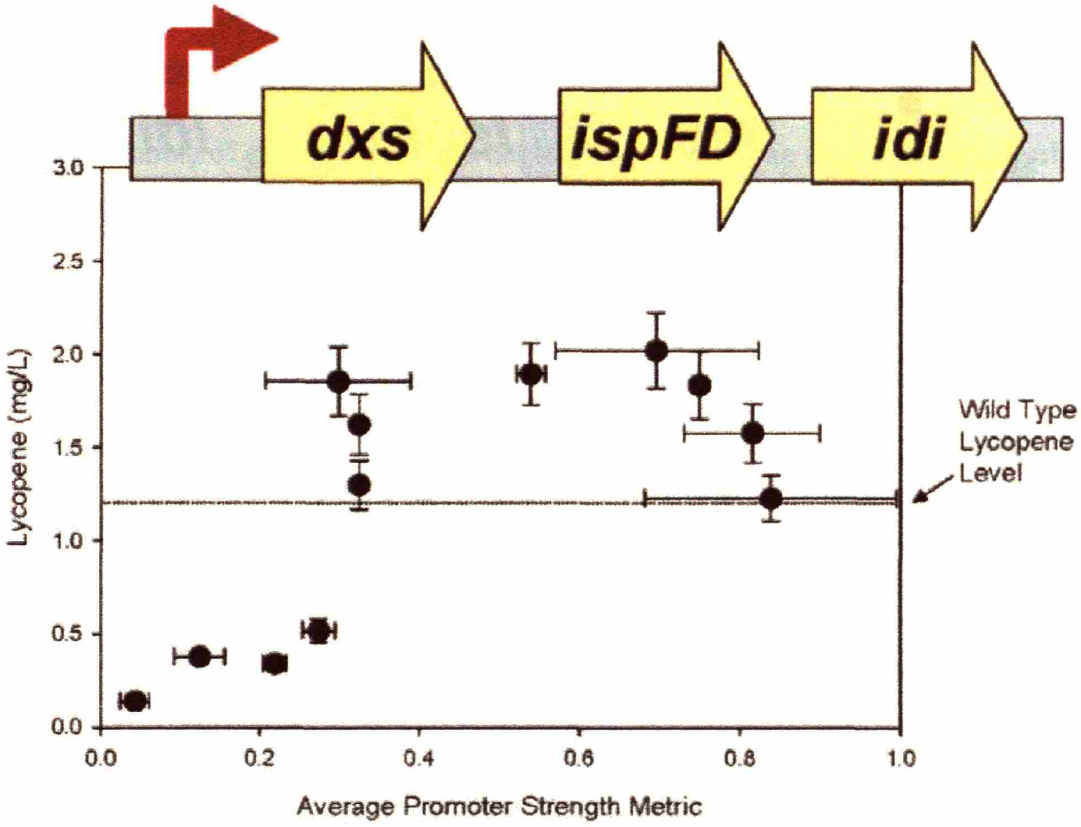

Figure 7.5: Lycopene yield and *dxs* activity in wild-type K12. Selected promoters were integrated in front of the *dxs* gene in a recombinant wild-type strain of *E. coli* and strains were later assayed for the production of lycopene. A clear maximum in lycopene production was obtained. From the wild-type production level, the native *dxs* promoter strength can be inferred to be between 0.2 to 0.4 according to our metric.

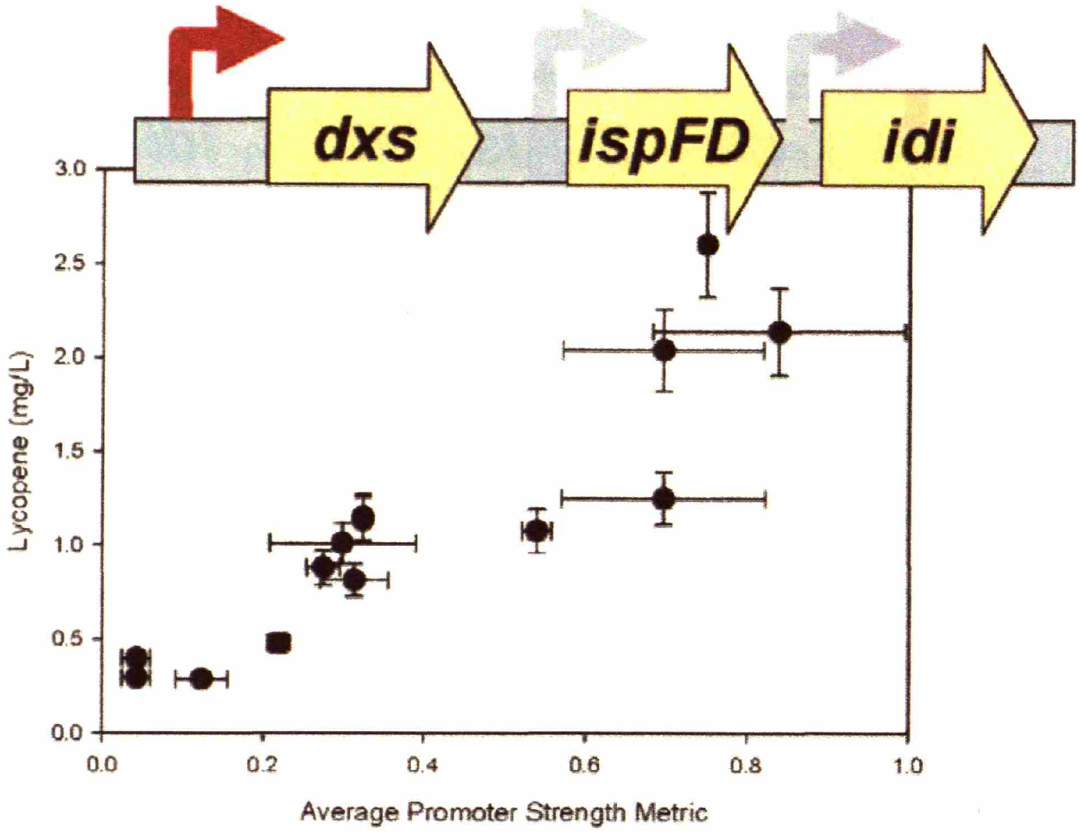

Figure 7.6: Lycopene yield and *dxs* activity in a pre-engineered strain. Selected promoters were integrated in front of the *dxs* gene in a recombinant strain also overexpressing *ispFD* and *idi*. In this case, the linear response of lycopene yield to the promoter strength illustrates a rate limiting behavior of *dxs* across all tested promoter strengths.

7.4.3 Lycopene yield and ppc activity

The tuning of gene expression is not limited to increasing the expression level of a given gene above the native expression level, but can also be extended to include the condition of gene knock-downs. As an example, **Section 4.3.1** indicated that *ppc* was identified through the stoichiometric model as a gene knockout target predicted to increase lycopene yield. However, a *ppc* knockout is lethal when grown in a medium with glucose as a sole carbon source. To investigate the impact of *ppc* activity on lycopene yield in a minimal medium without supplementation, three of the lowest strength promoters were integrated into the genome in front of the *ppc* gene. **Figure 7.7** illustrates the impact of *ppc* expression level on lycopene yield and juxtaposes the lycopene yield for the wild-type strain with a native expression level. Lycopene yields are plotted at the point of glucose exhaustion, which varied for each of the constructs. Furthermore, the cell yield was decreased with decreasing activity of *ppc*. In general, these results show the benefit of a *ppc* gene knockout when assaying solely the final lycopene yield from glucose. Nevertheless, these results illustrate the potential of the promoter library to perform gene knockdowns for essential genes in *E. coli*.

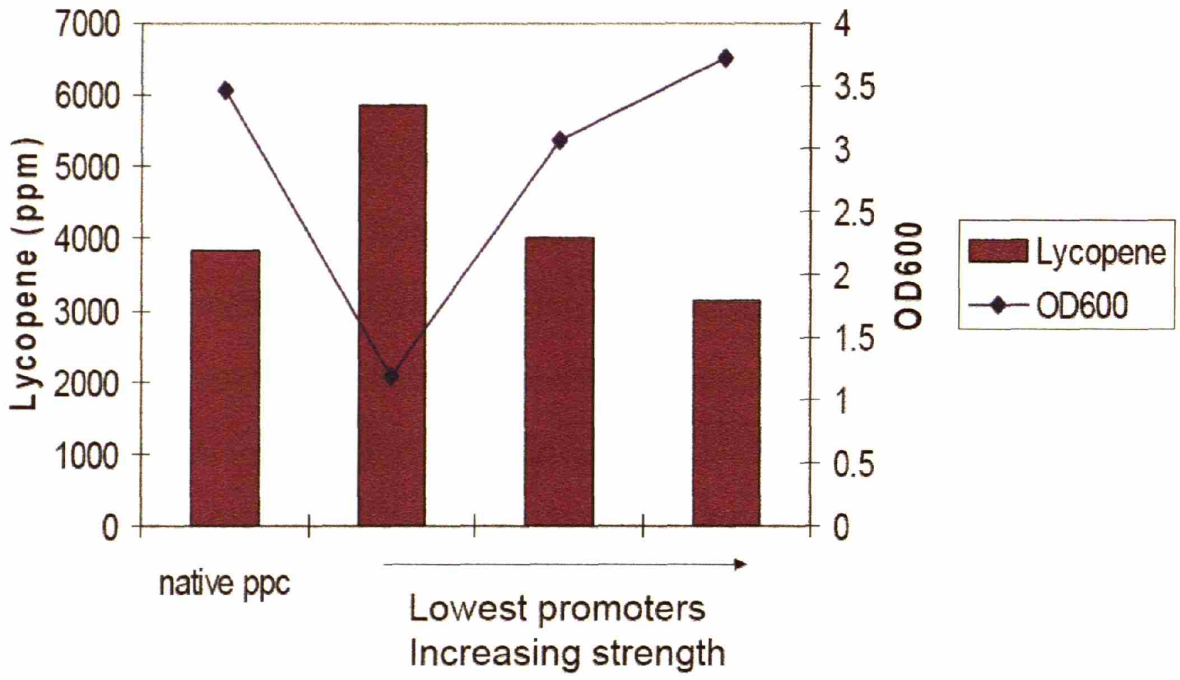

Figure 7.7: Lycopene yield and *ppc* activity. Selected low promoters were integrated in front of the *ppc* gene in a recombinant strain also overexpressing *ispFD* and *idi*. In this case, lycopene yield at the exhaustion of glucose is higher in the *ppc* knockdown strains, however, the growth yield is seen to decrease. These results highlight *ppc* as a legitimate gene knockout target for lycopene, however, also illustrates the lethality of a *ppc* knockout in glucose medium.

7.5 Summary

The nearly 200 random promoter mutants we screened varied widely in their expression strength and clonal expression heterogeneity. Screening for only those promoters which drive stable, monovariate expression in culture by flow cytometry was critical for deployment of our promoter constructs in pathway analysis and expression optimization. Isolating only the homogeneous promoters allowed us to establish a well-defined metric of promoter strength which combined data from several experimental assessments of gene expression levels. Using only a single technique to assess promoter strength often resulted in a scattering of the data, confounding the analysis of gene expression studies. The reliance on bulk averages would obscure the underlying relationship between expression and phenotype. The use of an integrated system allowed us to bypass the instabilities and inherent mutation rates associated with the over-expression of endogenous genes using plasmid-based systems (Zaslaver et al., 2004). Furthermore, this and other promoter libraries appear to have a broad host range (Jensen & Hammer, 1998), perhaps due to construction based on a heterologous constitutive promoter and reliance on the general polymerase machinery in the cell. This is exemplified through the three different strain backgrounds used in this study.

It is also possible to extend the promoter engineering concept to *Saccharomyces cerevisiae* as well. By screening a library of *TEF1* promoter mutants, also created by error-prone PCR, a promoter collection was obtained which drove a wide dynamic range of YFP production in *S. cerevisiae*. The creation of a library of promoter mutants in

yeast illustrates the applicability of this approach in both prokaryotic and eukaryotic contexts. As with *E. coli*, flow cytometry allowed isolation of only those promoters with relatively homogeneous reporter gene expression. Thus, the promoter engineering paradigm can yield libraries of promoter for precise genetic control despite the profound differences in bacterial and eukaryotic transcription mechanisms (Browning & Busby, 2004; T. I. Lee & Young, 2000).

Additionally, the analysis of libraries of promoters may be studied to deduce a linkage between sequence and phenotype. To this end, it would be possible to create correlations between mutation sites and promoter metrics such as strengths or variability in gene expression (Blake et al., 2003). Further application and study of this promoter library can greatly facilitate efforts in synthetic biology aiming to create synthetic genetic operons. The cataloging of promoter sequences along with their behavior can help in the selection of components to be used in synthetic gene networks such as toggle switches (Gardner, Cantor, & Collins, 2000) and for creating polygenic operons with prescribed ratios of gene expression.

For the first time, this work has created a general framework for the precise, *quantitative* control of gene expression *in vivo*. Our strategy allows (1) achievement of any desired expression level for a specific gene, (2) optimization of gene expression for maximal (or minimal) pathway function, and (3) a means for the analysis of the distribution of genetic control on pathway behavior. In two disparate examples we have shown that pathway function can exhibit well-defined extrema with respect to levels of gene expression. The existence of these extrema evinces the need for precise gene-dosage studies for the full understanding of pathway behavior. The creation and detailed

characterization of a promoter library as described here is a facile and robust means to such an end. Furthermore, it is possible to utilize these promoters to investigate and optimize cellular systems and probe into the metabolic landscape. However, developing a method for identifying which genes need to be optimized is still an open question. To address this issue on a global scale, the next chapter describes a method for altering multiple genes simultaneously.

Chapter 8

global Transcription Machinery Engineering (gTME)

8.1 Motivation

It is now generally accepted that important cellular phenotypes are affected by many genes. As a result, engineering or enhancing our understanding of a desired phenotype would be facilitated enormously by simultaneous multiple gene modification. Furthermore, limitations in the capacity to traverse metabolic landscapes through single gene perturbations linked with a search strategy justify the need for a novel approach to whole-cell engineering. However, the capacity to introduce such modifications has remained an elusive task for cellular and metabolic engineering. This chapter presents a method that allows modulation of multiple gene expression at the highest level with profound implications for phenotype improvement of prokaryotic and eukaryotic cells alike.

Cellular systems have optimized the capacity to self-regulate their thousands of genes through fine-tuning components of global transcription machinery. In bacterial systems, sigma factors focus the promoter preferences of the RNA polymerase (Burgess

& Anthony, 2001) and preliminary molecular biology evidence indicates that mutations to key residues can alter this preference (Gardella, Moyle, & Susskind, 1989; Malhotra, Severinova, & Darst, 1996; Owens et al., 1998; D.A. Siegele et al., 1989). First, this chapter will demonstrate that these components of global cellular transcription machinery can be engineered to elicit complex phenotypes controlled by multiple genes. This novel approach allows the high throughput probing of a vastly unexplored search space by evaluating multiple, simultaneous gene alterations. As a part of proof-of-concept, this tool of global Transcription Machinery Engineering (gTME) will be used to investigate numerous, distinct phenotypes in *E. coli.* In each case, the tool of global Transcription Machinery Engineering (gTME) outperformed traditional approaches, exceeding, *in a matter of weeks*, benchmarks achieved through decades of research. Through gTME, it is now possible to unlock complex phenotypes regulated by multiple genes which would be very unlikely to reach by the relatively inefficient, iterative gene-by-gene search strategies. Finally, the generic nature of this approach is exploited through examples in eukaryotic systems (yeast) as well.

8.2 Background

Multiple genetic modifications are necessary, in general, to unlock latent cellular potential. However, most current cellular and metabolic engineering approaches rely almost exclusively on the deletion or over-expression of single genes due to experimental limitations in vector construction, transformation efficiencies, and screening capacity. These limitations preclude the simultaneous exploration of multiple gene modifications

and confine gene modification searches to restricted sequential approaches where a single gene is modified at a time. As a result, current paradigms relying predominantly on these limited types of modifications, often fail to reach a global phenotype optimum due to the complexity of metabolic landscapes (Alper et al., 2005b; Alper, Miyaoku, & Stephanopoulos, 2005) and inability of incremental or greedy search algorithms to uncover mutants that are beneficial only when multiple modifications are simultaneously introduced. To address these limitations, alternative methods have been investigated, however, these approaches are often inherently limited in scope and focus due to the reliance on specific transcription factors or DNA binding motifs (Gerber et al., 1994; J. S. Kim et al., 1997; Park et al., 2003).

The modification and engineering of the global transcription machinery presented here provides the means to making higher-level modifications which can traverse transcriptional control schemes and diverse pathways. As such, modified transcription machinery units offer the unique opportunity to introduce *simultaneous global transcription-level alterations* that have the potential to impact cellular properties in a very profound way. This tool exploits the global regulatory functions of the bacterial σ^{70} sigma factor to introduce multiple simultaneous gene expression changes and thus facilitate whole-cell engineering by selecting mutants responsible for a variety of improved cellular phenotypes.

8.3 Implementation

The main sigma factor, σ^{70}, was subjected to random mutagenesis and introduced into *E. coli* to search for varying cellular phenotypes. This sigma factor was chosen on the premise that mutations will alter the promoter preferences of RNA polymerase affecting transcription rates and thus modulating the transcriptome at a global level (Gardella, Moyle, & Susskind, 1989; Malhotra, Severinova, & Darst, 1996; Owens et al., 1998; D.A. Siegele et al., 1989). The *rpoD* gene and native promoter region were subjected to error-prone PCR and cloned into a low-copy expression vector. A nearly 10^6 viable-mutant library was initially constructed and transformed into strains. **Figure 8.1** depicts the basic methodology for gTME.

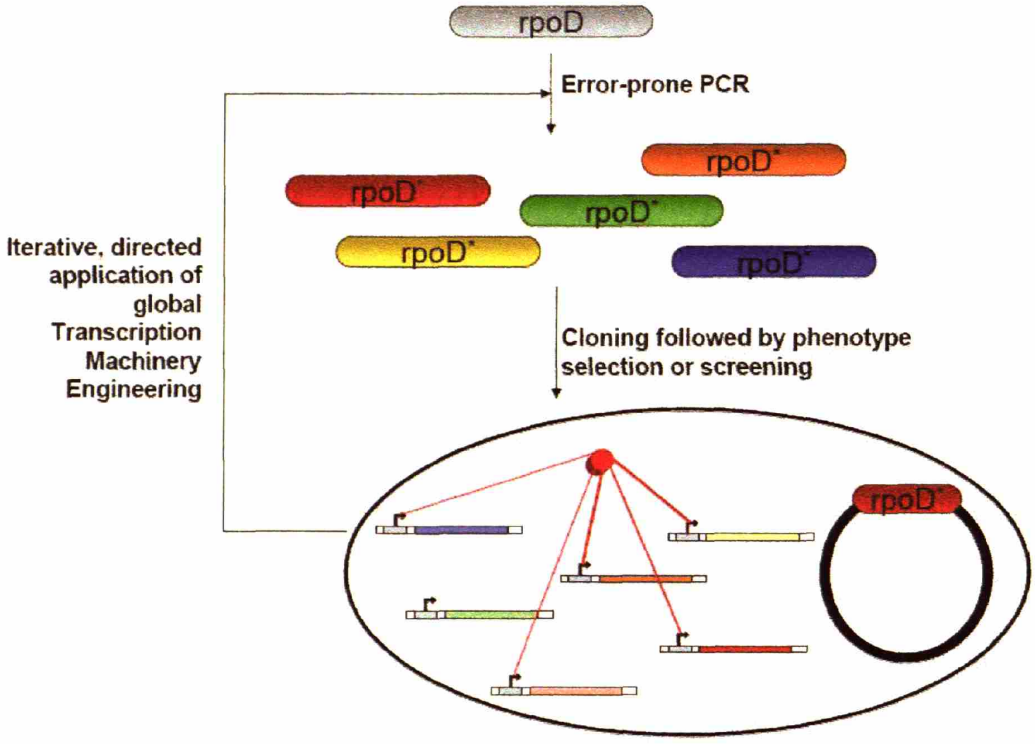

Figure 8.1: Basic methodology of global transcription machinery engineering. By introducing altered global transcription machinery into a cell, the transcriptome is altered and the expression level of genes changes in a global manner. As an example, the bacterial sigma factor 70 (encoded by *rpoD*) was subjected to error-prone PCR to generate various mutants. The mutants were then cloned into a low-copy expression vector, during which the possibility arose for a truncated form of the sigma factor due to the presence of a nearly complete internal restriction enzyme site. The vectors were then transformed into *E. coli* and screened based on the desired phenotype. Isolated mutants can be subjected to subsequent rounds of mutagenesis and selection to further improve phenotypes.

8.4 *E. coli* Applications

Many distinct and diverse phenotypes of (1) tolerance to ethanol, (2) metabolite overproduction, (3) multiple, simultaneous phenotypes, (4) tolerance to acetate, (5) tolerance to pHBA, and (6) tolerance to hexane were investigated as proof-of-concept. Most of these phenotypes has been studied by traditional methods of randomized cellular mutagenesis, gene complementation and knockout searches, and microarray analysis, with limited success to-date (Gill et al., 2002; Gonzalez et al., 2003; Hemmi et al., 1998; Zaldivar, Nielsen, & Olsson, 2001).

8.4.1 Ethanol tolerance

Mutants of the sigma factor library were first selected on the basis of ability to grow in the presence of high concentrations of ethanol in complex medium (Yomano, York, & Ingram, 1998), a phenotype which, at present, is limiting prospects of industrial bioethanol production (Zaldivar, Nielsen, & Olsson, 2001). For this selection, strains were serially subcultured twice at 50 g/L of ethanol overnight, then plated to select for tolerant mutants. A total of 20 colonies were selected and assayed for growth. After confirming that the improved phenotype was conferred by the mutant factor, the best mutant sigma factor was subjected to two additional rounds of mutation and selection. With both subsequent rounds, the selection concentration was increased to 60 and 70 g/L of ethanol. In these enrichment experiments, mutants were isolated after 4 and 8 hours of incubation due to the strong selection pressure used. Isolated mutants from each round show improved overall growth at all ethanol concentrations tested (**Figure 8.2**).

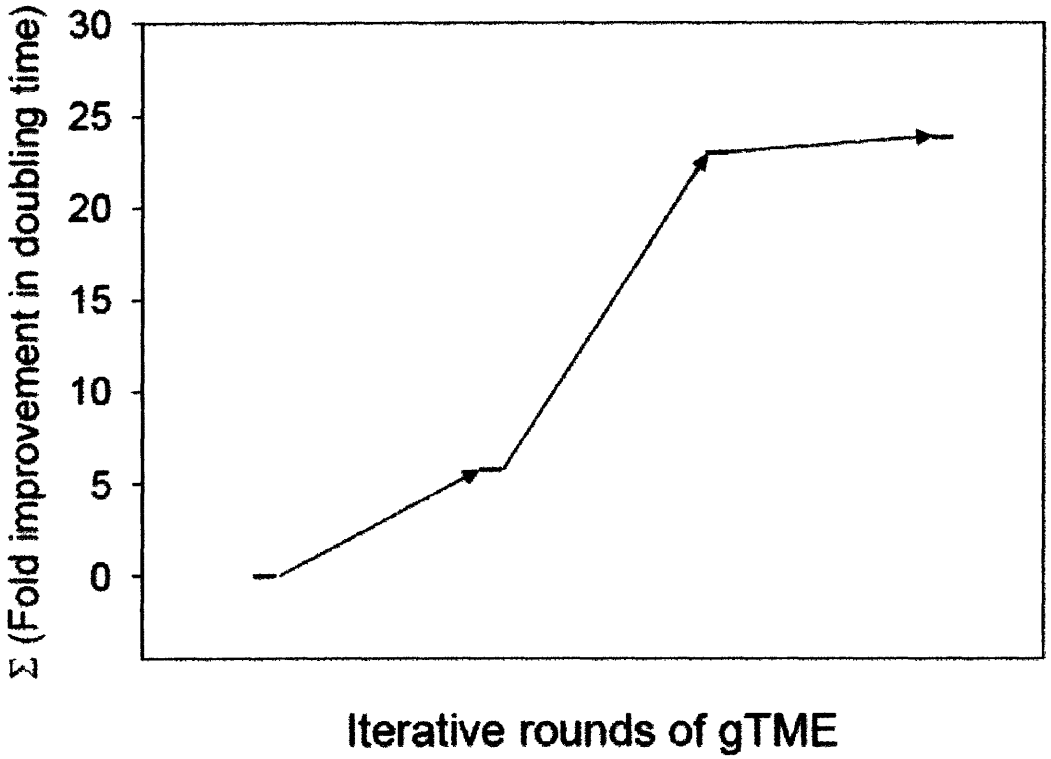

Figure 8.2: Overall improvement of ethanol tolerance using gTME. The overall enhancement of phenotype through the various rounds of mutagenesis to identify mutant factors. Overall enhancement (y-axis) is assessed by taking the summation of the fold reduction of doubling time for the mutant over the control at 0, 20, 40, 50, 60, 70 and 80 g/L of ethanol. By the third round, the improvement in growth rate seems to be incremental.

Figure 8.3 identifies the sequences of the best mutants isolated from each round of mutagenesis. Interestingly, the second round mutation led to the formation of a truncated factor which is apparently instrumental in increasing overall ethanol fitness. This truncation, arising from an artifact in the restriction enzyme digestion and sequence similarities, includes part of region 3 and the complete region 4 of the protein. Region 4 is known to be responsible for binding to the promoter region and to anti-sigma factors, and a truncated form has been previously shown to have an increased binding affinity relative to that of the full protein (Sharma et al., 1999). It is therefore possible that this truncated mutant serves to act as a potent and specific inhibitor of transcription by binding to preferred promoter regions (or anti-sigma factors) and preventing transcription since the protein region responsible for recruiting the polymerase is absent in the truncation. Furthermore, mutations in the R603 site, which occurred in rounds 1 and 2, have been implicated in reduction in transcriptional capacity at most promoters tested (Lonetto et al., 1998). In the truncated form of the round 2 mutant, the I511V mutation of the first round was reverted back to an isoleucine, leaving only one mutation. Finally, the mutant identified in the third round was a truncated factor with 8 additional mutations. These rounds of mutagenesis and resulting sequences suggest an important distinction compared with protein directed evolution. In the latter case, mutations which increase protein function are typically additive in nature (Wells, 1990; Zhang et al., 1995). However, this is certainly not the case when altering transcription machinery as these factors act as conduits to the transcriptome. In this regard, many local maxima may occur in the sequence space due to the various subsets of gene alterations which may lead to an improved phenotype.

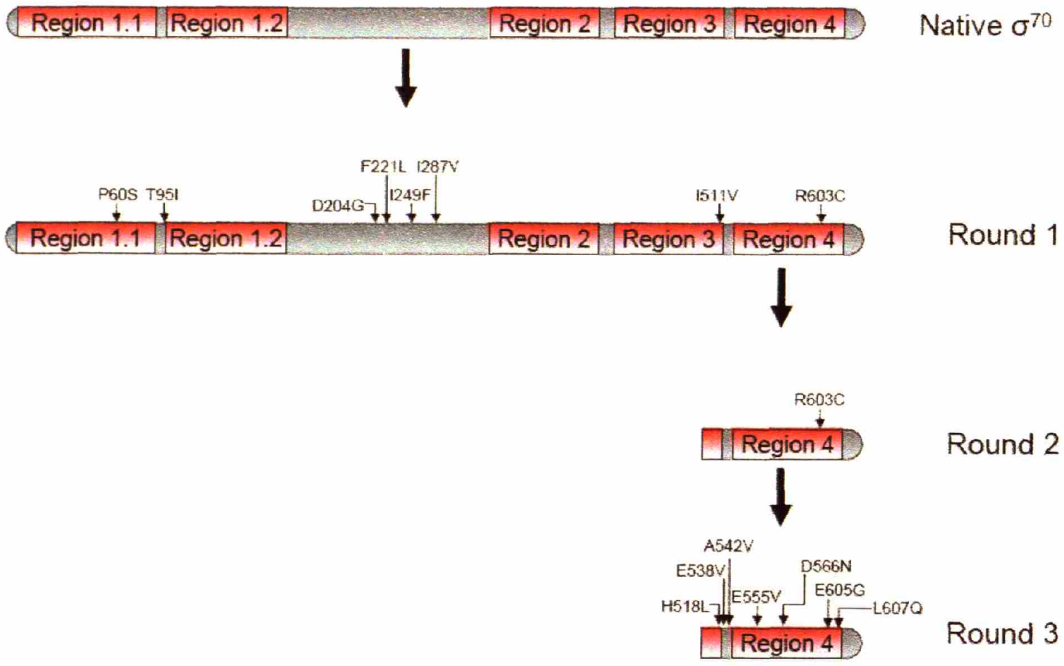

Figure 8.3: Sequence analysis of ethanol sigma factor mutants. The location of mutations on the σ^{70} protein are indicated in relation to previously identified critical functional regions (Gruber & Gross, 2003). The second round mutagenesis resulted in the identification of a truncated factor containing only one of the two prior mutations in that region.

All isolated strains harboring the mutant sigma factors exhibited increased growth rates and overall tolerance relative to the control at elevated ethanol concentrations. Interestingly, the growth phenotype of the mutant strains in the absence of ethanol was not impacted (**Table 8.1**). With each subsequent round of mutagenesis, cells were able to sustain growth for longer than 8 hours at ever-increasing ethanol concentration before succumbing to the ethanol toxicity, marked by a decrease in cell density. The pronounced increase in ethanol tolerance obtained through this method is illustrated by the growth curves of the round 3 strain shown in (**Figure 8.4**) along with those of the wild type sigma factor control.

	Doubling Time (h)	Ratio of doubling times ($t_{d,control}$ / $t_{d,engineered\ mutant}$)		
Ethanol Concentration (g/L)	Control	Round 1	Round 2	Round 3
0	0.76	1.01	0.98	0.98
20	1.31	1.68	1.63	1.63
40	2.41	1.64	1.30	1.54
50	7.24	1.92	1.82	2.06
60	69.3	4.53	11.70	11.18
70	192.3	1.40	11.56	12.43
80	ND	ND	28.64 hours	29.80 hours
Maximum sustainable concentration (g/L)	40	50	60	70

Table 8.1: Improvement of ethanol tolerance through engineered sigma factors. Improvements in the fold reduction of doubling time are presented for increasing concentrations of ethanol for the three rounds of directed evolution. Interestingly, the growth rate phenotype of these mutants in the absence of ethanol is not impacted, as the growth rate was the same as the control.

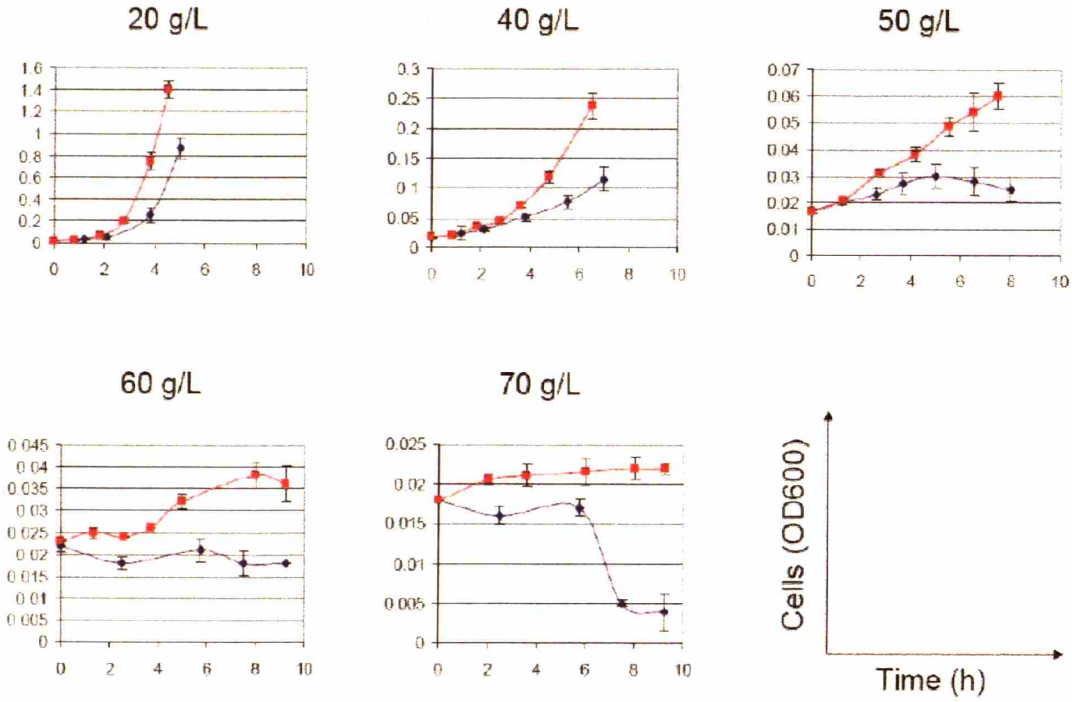

Figure 8.4: Growth curves for ethanol-tolerant sigma factor mutants. Growth curves are presented for the Round 3 mutant harboring the mutant σ^{70} (Red) and control strains harboring the wild type σ^{70} (Blue). The round 3 mutant has significantly improved growth rates at all tested ethanol concentrations.

8.4.1.1 Transcriptional analysis

In an effort to further elucidate the mechanism of action of the mutant sigma factors, the transcriptome of these strains was assayed using DNA microarrays under various conditions. First, all strains (including the control) were assayed in the absence of ethanol to assess the impact of the mutant sigma factors on transcription in normal medium. In general, the transcriptional results validated the capacity of mutant sigma factors to elicit simultaneous global transcription-level alterations. Thus, a total of 72 genes were differentially expressed in cells harboring the third round mutant compared to the control at a p-value threshold of 0.001 (44 of these genes were upregulated, with the remaining 28 genes downregulated). A total of 125 genes are changed with the first round mutant and 82 are changed with the second round mutant. These results suggest that mutant sigma factors through each round are converging on a subset of important genes, despite the deviations seen in the sequence data. Furthermore, they results echo prior observations suggesting that ethanol tolerance is a phenotype controlled by many genes (Gonzalez et al., 2003).

Next, the transcriptomes of the best mutant from round 3 and the control strain were further assayed at varying levels of ethanol (20 and 40 g/L for round 3 and 20 g/L for the control). Collectively, these gene expression profiles provide the basis for an initial understanding of the underlying mechanism of enhanced ethanol tolerance supported by the mutant factors. **Figure 8.5** illustrates the complex, pleiotropic impact of ethanol in the control strain. Ethanol initiated a generic stress response consisting of 354 genes differentially expressed (at a p-value threshold of 0.001), many of which are typically associated with cellular stress responses. The mutant sigma factor (in the

absence of ethanol) alters significantly fewer, but still a good number of genes, some of which overlap with the generic stress response (**Figure 8.6**). However, this strain is able to grow in the presence of elevated ethanol and similarly, the response to ethanol is varied compared with the control (**Figure 8.7**). In this new response, many genes previously related with ethanol stress response in the wild-type have now been pre-programmed by the sigma factor, while an additional set of genes are altered by ethanol, representing a new mode of response. It is interesting to note a substantial change in iron-related enzymes negatively controlled by the small RNA encoded by *ryhB* (Masse & Gottesman, 2002) (which is upregulated by the sigma factor and by ethanol in the control). In particular, there is a substantial decrease in expression of genes for ferritin (*ftnA*) as well as metabolic enzymes containing iron-based catalytic domains (*sdhABCD*, *acnA*). Overall, the response to ethanol is tempered by the mutant factor. While a total of 354 genes change in the control when treated with 20 g/L ethanol, only 117 genes comprise the ethanologenic response in the mutant strain above the genes altered by the sigma factors. Based on these analyses, it appears that the mutant sigma factors temper the transcriptional response compared with the control, leading to the increased ethanol tolerance. The expression level changes in these genes are summarized in **Table 8.2 through Table 8.4**

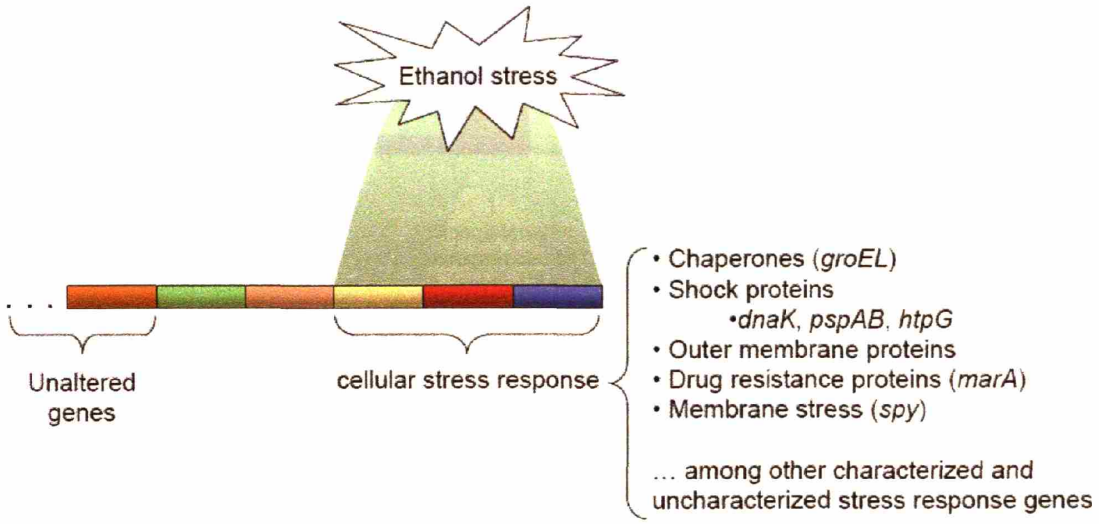

Figure 8.5: Transcriptional analysis of general ethanol stress. The control strain was tested and compared two conditions: 0 and 20 g/L ethanol. Schematics are presented which represent the mode of action of ethanol response. Blocks represent groups of genes which are differentially expressed (either up or down regulated) under similar conditions. A few genes of interest which are differentially expressed from each class are highlighted in the figure. Ethanol acts on the control strain by eliciting a general cellular stress response comprising 354 genes differentially expressed (p-value of 0.001 or less), including several well characterized stress response genes.

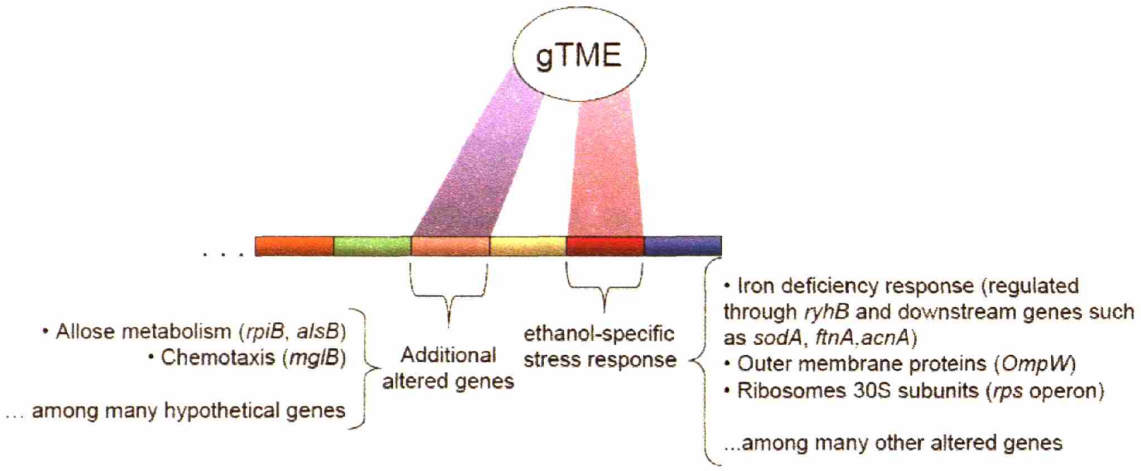

Figure 8.6: Transcriptional analysis of an ethanol sigma factor mutant. Schematics are presented which represent the mode of action of the mutant sigma factors (gTME). Blocks represent groups of genes which are differentially expressed (either up or down regulated) under similar conditions. A few genes of interest which are differentially expressed from each class are highlighted in the figure. The mutant sigma factors, in the absence of ethanol changes the expression of many genes, some of which overlap with the generic stress response, which suggest a subset of genes responsible for ethanol tolerance.

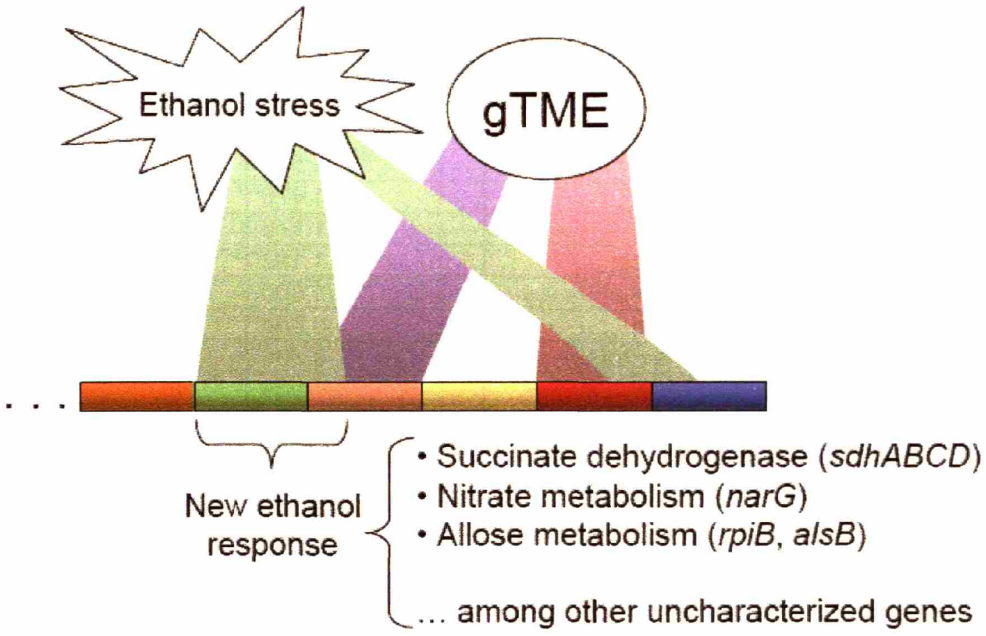

Figure 8.7: Transcriptional analysis of an ethanol sigma factor mutant in response to ethanol. Schematics are presented which represent the mode of action of the mutant sigma factors (gTME). Blocks represent groups of genes which are differentially expressed (either up or down regulated) under similar conditions. A few genes of interest which are differentially expressed from each class are highlighted in the figure. In the presence of ethanol, the mutant strain has a tempered, but varied response to ethanol. This response is less extreme than the control which correlates with the increased growth rates.

b-number	Gene Name	Function	log fold expression ratio[1]	p-value for differential expression
b4142	*groE*	chaperone	1.545	5.27E-03
b4143	*groL*	chaperone	1.476	2.23E-04
b0014	*dnaK*	shock protein	1.176	4.76E-04
b1304	*pspA*	phage shock protein	1.182	5.41E-04
b1305	*pspB*	phage shock protein	1.000	4.72E-03
b0473	*htpG*	heat shock protein	0.866	3.01E-04
b0929	*ompF*	outer membrane protein	-2.386	4.49E-06
b1531	*marA*	multiple drug response	1.034	1.96E-03
b1743	*spy*	membrane stress	1.891	6.63E-03

[1] Log-fold ratio comparing the control strain in the presence of 20 g/L ethanol to the control strain with 0 g/L ethanol, log ratio is expressed in base 2.

Table 8.2: Change in expression of ethanol response genes. Changes in gene expression for the genes discussed in **Figure 8.5** are summarized.

b-number	Gene Name	Function	log fold expression ratio[2]	p-value for differential expression
b4090	*rpiB*	ribose-5-phosphate isomerase / allose-6-phosphate isomerase	1.981	2.21E-05
b4088	*alsB*	allose binding protein	2.329	7.07E-06
b2150	*mglB*	chemotaxis gene	0.744	4.90E-04
b4451	*ryhB*	small regulatory RNA	0.793	5.33E-03
b3908	*sodA*	Superoxide dismutase	2.246	6.12E-05
b1905	*ftnA*	ferritin	-1.088	2.46E-03
b1276	*acnA*	aconitase	0.772	5.07E-04
b1256	*ompW*	outer membrane protein	-2.884	3.54E-05
b0023	*rpsT*	small ribosome	-0.677	5.19E-03
b3065	*rpsU*	small ribosome	-0.677	5.60E-03

[2] Log-fold ratio comparing the third round mutant at 0 g/L ethanol to the control strain with 0 g/L ethanol, log ratio is expressed in base 2.

Table 8.3: Change in expression of sigma factor mutant-induced genes. Changes in gene expression for the genes discussed in **Figure 8.6** are summarized.

b-number	Gene Name	Function	log fold expression ratio (mutant)[3]	p-value for differential expression	log fold for control in ethanol[1]	p-value for differential expression
b0723	*sdhA*	Succinate dehydrogenase	-1.685	7.71E-04	0.401	2.32E-01
b0724	*sdhB*	Succinate dehydrogenase	-1.329	2.41E-04	0.703	1.04E-02
b0721	*sdhC*	Succinate dehydrogenase	-2.069	2.27E-04	-0.185	3.12E-01
b0722	*sdhD*	Succinate dehydrogenase	-1.734	2.01E-05	0.053	8.63E-01
b1224	*narG*	nitrate metabolism	-1.376	1.22E-03	0.023	8.89E-01
b4090	*rpiB*	ribose-5-phosphate isomerase / allose-6-phosphate isomerase	-2.203	8.86E-05	-0.004	9.61E-01
b4088	*alsB*	allose binding protein	-2.262	9.65E-05	0.177	1.85E-02

[1] Log-fold ratio comparing the control strain in the presence of 20 g/L ethanol to the control strain with 0 g/L ethanol, log ratio is expressed in base 2.

[3] Log-fold ratio comparing the third round mutant at 40 g/L ethanol to the third round mutant with 0 g/L ethanol, log ratio is expressed in base 2.

Table 8.4: Change in expression of new ethanol response. Changes in gene expression for the genes discussed in **Figure 8.7** are summarized.

The transcriptional analysis also yielded several interesting types of gene expression patterns which may provide access points to key genes in the mechanism of ethanol tolerance caused by the mutant sigma factors. **Figure 8.8** presents three representative, yet varied patterns of gene expression present in these strains. Pattern 1 comprises genes which show a dose response to ethanol. Their expression reaches the same value of the control, but only at higher ethanol concentrations and highlights the tempered response of the mutants to ethanol. Pattern 2 presents an interesting response where the mutant sigma factors impact a gene in a significant manner and the ethanol acts to reverse this effect. These genes allow for some buffering of gene expression which could lessen the impact of ethanol in these strains. Finally, pattern 3 represents several genes which are altered by the sigma factors in the same manner that ethanol alters the strain. In a way, these genes represent a priming of the cell to respond to ethanol more aptly. It is evident from these analyses that ethanol tolerance is highly pleiotropic and regulated by a multitude of genes. The putative targets extracted from this analysis can provide invaluable leads to key genes responsible for ethanol tolerance.

These results, (i) illustrate that gTME was able to increase the ethanol tolerance beyond the levels previously reported in the literature using more traditional methods, (ii) highlight the application of sequential rounds of refinement for further improving the cellular phenotype, and (iii) illustrate the importance of making multiple, simultaneous alterations of genes expression to obtain phenotypes of interest. At least three distinct and important modes-of-action: (1) priming, (2) buffering, and (3) tempering of the transcriptome response to ethanol arose out of the transcriptional analysis. Each of these may contribute to different degrees to the overall phenotype improvement.

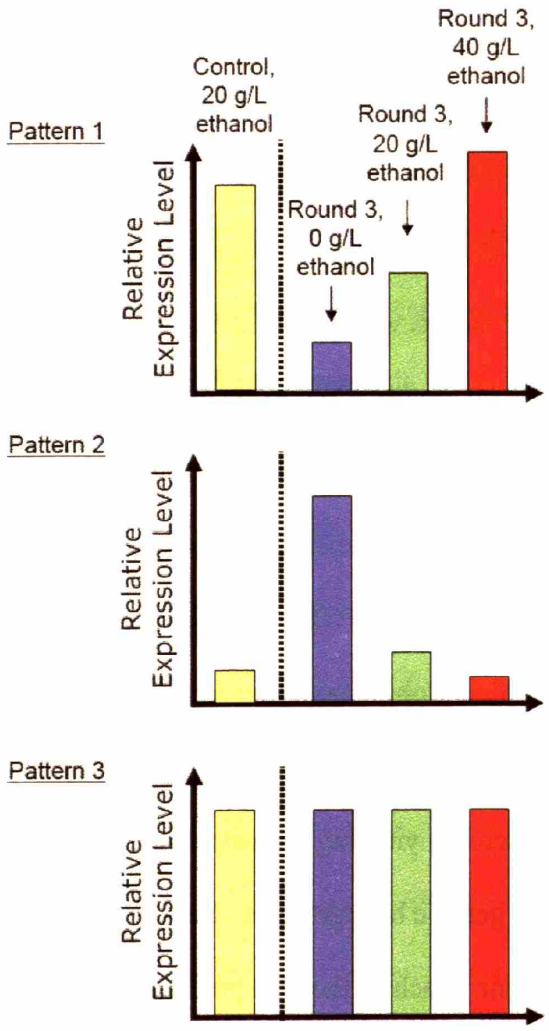

Figure 8.8: Patterns in the transcriptional profiles in response to ethanol. Several patterns of gene expression emerge from the analysis of the microarray data. In pattern 1, several genes in the mutant strain show a slower, dose-response to ethanol compared with the control. In pattern 2, several genes are altered by the sigma factor and reversed in the presence of ethanol, allowing for a buffering of the expression. In pattern 3, several genes are changed by the sigma factors in a way which is similar to the ethanol response, priming the cell for ethanol.

8.4.2 Lycopene Production

The method of gTME was found to be effective for improving the phenotype of metabolite overproduction. Previously, the use of systematic and combinatorial gene knockout searches were used to identify genetic targets which enhanced lycopene production in the background of a pre-engineereed strain (Alper et al., 2005b; Alper, Miyaoku, & Stephanopoulos, 2005). Here, the technique of gTME was explored to enhance lycopene production and compare the impact to traditional metabolic engineering approaches. Utilizing the parental strain (K12 PT5-*dxs*, PT5-*idi*, PT5-*ispFD* harboring pAC-LYC), Δ*hnr*, and the two identified global maximum strains, Δ*gdhA*Δ*aceE*Δ*fdhF*, and Δ*gdhA*Δ*aceE*Δ$_{P}$*yjiD*, it was possible to search for and identify mutant sigma factors (based on a colorimetric screen) yielding increased lycopene production, independently in each of the above genetic backgrounds. Several mutants were chosen based on increased lycopene content. Each of the best producing mutants from these selected strains harbored different mutated versions of a truncated *rpoD*, although several, suboptimal, whole-length mutants were also recovered. Sequences for these mutants are provided in **Figure 8.9**.

Figure 8.10 illustrates the lycopene content after 15 hours for several strains of interest. The single round of gTME in both the parental strain and *hnr* knockout was able to achieve similar increases in lycopene accumulation as strains previously engineered through the introduction of three distinct gene knockouts. Furthermore, in the backgrounds of these knock out mutants, lycopene levels were further increased through the introduction of an additional, yet distinct mutant sigma factor. These results suggest

that, (i) gTME is able to elicit phenotypes of metabolite overproduction and, more importantly, (ii) a single round of selection using gTME is more effective than several rounds of typical gene knockout or overexpression modifications linked with a search strategy. Moreover, comparing the results of **Figure 8.10**, it is clear that gTME is able to not only enhance but surpass the capabilities of the knockout strains obtained from selection using traditional methods.

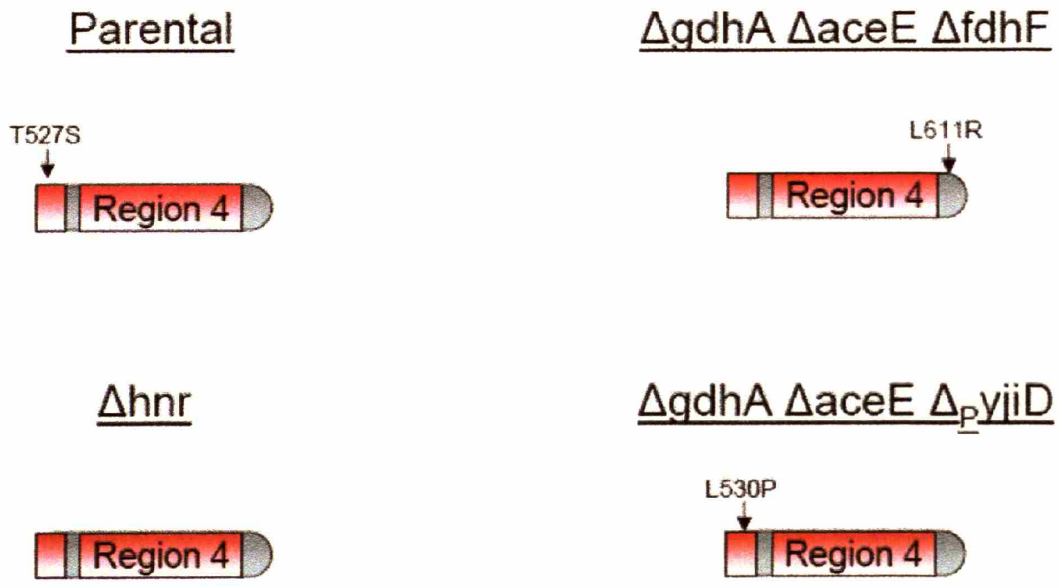

Figure 8.9: Sequences for lycopene sigma factor mutants. Schematics for the identified mutants increasing lycopene production are provided. While each of the identified mutants was truncated, each possessed a unique set of mutations. Furthermore, the mutant identified from the *hnr* knockout background was simply truncated and contained no mutations.

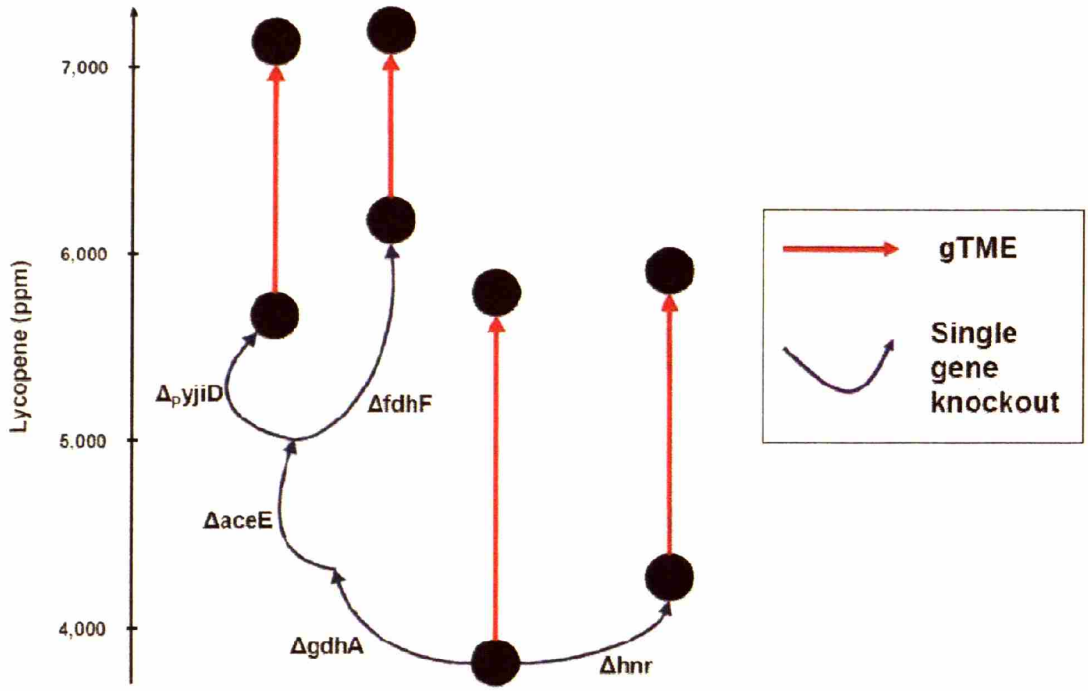

Figure 8.10: Application of gTME to a metabolite production phenotype. The tool of gTME was compared with traditional methods of strain improvement whereby rational and combinatorial methods are applied to the identification of gene knock out targets aiming to enhance lycopene accumulation in an engineered strain of *E. coli*. The mutant sigma factor library was introduced into four pre-engineered lycopene over-producing strains to identify factors which further increase production. Lycopene content, in ug/g dry cell weight (ppm), are presented after 15 hour cultivations. These results indicate that gTME is a powerful tool for eliciting metabolic phenotypes and, more importantly, is more effective than the single gene knockout or overexpression iterations.

The four strains with varying genetic backgrounds were then combined with the four independently identified mutant sigma factors to examine the resulting 16 strain landscape. It is interesting to initially note that none of the identified mutants which were sequenced for a given genetic background overlapped with those identified in another genetic background. As a result, it is initially suspected that the landscape would be diagonally dominant, indicating that the effect elicited by the mutant factor is specific to the genetic background. These 16 strains along with the controls were cultured in a 2x M9 medium with staged glucose feed. The lycopene level was assayed at 15, 24, 39, and 48 hour timepoints. **Figure 8.11** presents a dot plot which depicts the maximum fold increase in lycopene production achieved over the control during the fermentation. The size of the circle is proportional to the fold increase. As suspected, the landscape is clearly diagonally-dominant with mutant factors predominantly working in the strain background in which they were identified. These results suggest that different transcriptome reprogramming is required for lycopene production in different genotypes. As an example of these modes, the maximum fold difference in the wild type strain was realized after only 15 hours and then converged with the control strain by the end of the fermentation. Conversely, the mutant factor in the Δ*gdhA*Δ*aceE*Δ$_{P}$*yjiD* strain progressively increased in lycopene content compared with the control for increasing timepoints. When limited to only on round of gTME selection, the highest lycopene production resulted from using genetic backgrounds of a previously engineered strain. However, the results of ethanol tolerance suggest that it is possible to achieve continual improvements in fitness through the application of multiple rounds of evolution,

indicating that it may be possible to increase lycopene production further. From this analysis, it appears that, in general, the mutant sigma factors were not transferable across strain backgrounds, which suggests that the required mode of transcriptional reprogramming is genotype-specific.

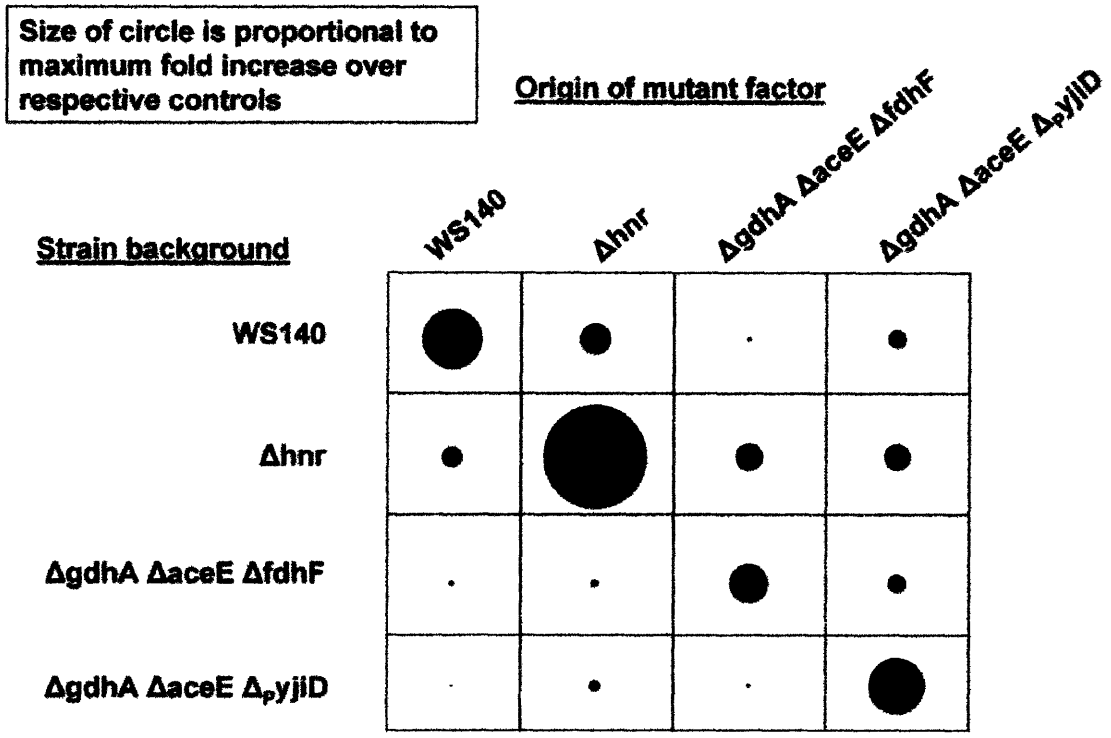

Figure 8.11: Genotype specificity of identified sigma factor mutants. The four strains with varying genetic backgrounds were then combined with the four independently identified mutant sigma factors to examine the resulting 16 strain landscape. This dot plot, representing the maximum fold increase in lycopene, indicates that, in general, mutant sigma factors are not transferable between genotypes suggesting that the transcriptional reprogramming is genotype-specific.

8.4.3 Multiple tolerances

The tool of gTME was studied as a method to impart *multiple, simultaneous* phenotypes to a cell. For this study, the multiple tolerance phenotype of ethanol and sodium dodecyl sulfate (SDS) was chosen. As a multiple phenotype may be elicited through several different trajectories, sigma factor mutants were isolated by following four distinct search strategies (**Figure 8.12**): (1) isolate first an ethanol tolerant mutant, then create a new mutant library and screen for ethanol and SDS tolerance; (2) isolate first an SDS mutant, then create a new mutant library and screen for ethanol and SDS tolerance; (3) select for an ethanol/SDS tolerant mutant simultaneously on the single library; or, (4) independently select for an ethanol and an SDS mutant and then co-express these two proteins. The best mutant strains obtained through each approach were assayed using as metric their growth rate under all possible combinations of 0, 0.5, 1% w/w SDS and 0, 25, 50 g/L ethanol. Total fitness (see **Fig. 8.12** for definition) is a measure of the extent to which the mutant is able to outperform the control under all nine possible conditions. On the other hand, ethanol fitness and SDS fitness represent the growth enhancement when only one of the two toxic compounds is varying while the other is kept at the control level of 0 g/L. **Figure 8.12** summarizes the results of the four possible search strategies. In both the sequential searches and the simultaneous search (strategies 1-3), there exists a tradeoff between total fitness and pure component fitness (either SDS or ethanol or both). Of these three routes, the sequential path of selecting for ethanol first, followed by a new mutagenesis step and selection in ethanol/SDS is superior. However, the co-expression of the full-length ethanol mutant and the truncated SDS mutant imparted the most significant phenotype (highest overall fitness) without a

sacrifice of pure component fitness, which was present in all the remaining search strategies. In a way, co-expression effectively allowed the additive expression of the two independently identified phenotypes. The strain with co-expressed mutants had a similar individual component fitness compared with the single-phenotype mutant (0.87 vs. 0.89 for ethanol and 0.15 vs. 0.18 for SDS) along with a greatly improved total fitness. In particular, no single mutant factor was identified which could impart fitness comparable to that of co-expression. As such, the expression of a full length and truncated mutant could be a potent method for directing overexpression and knockout modifications simultaneously in the cell. These results suggest a powerful method for creating desirable phenotypes. Sequences for these mutants are provided in **Figure 8.13**.

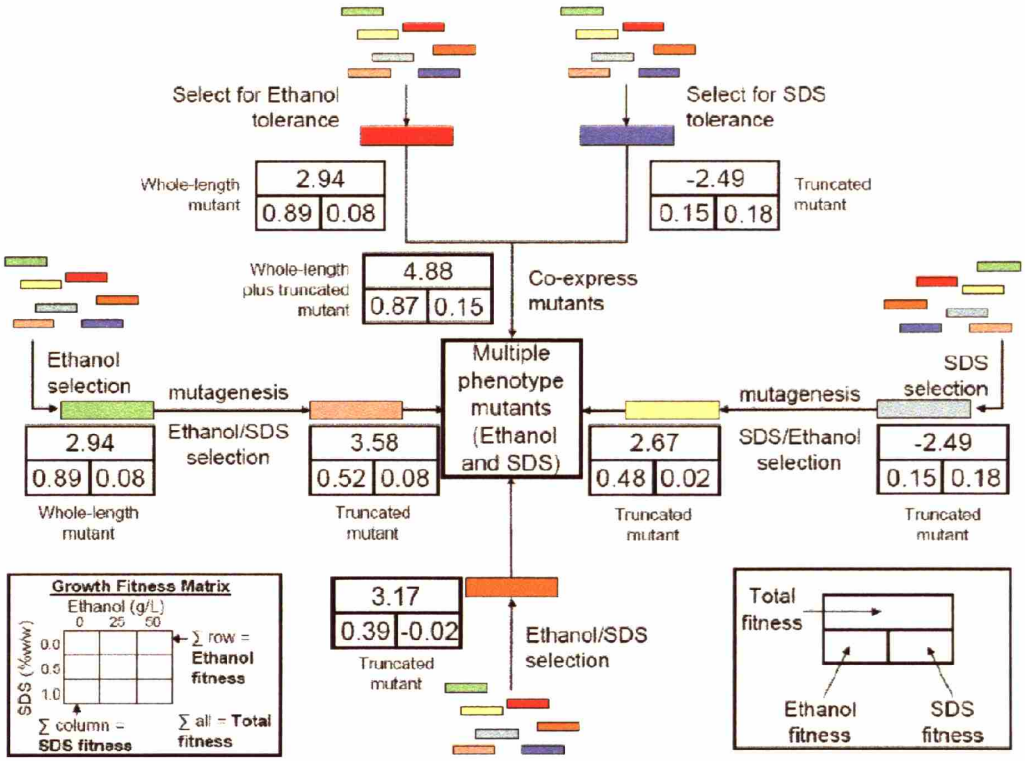

Figure 8.12: Eliciting multiple, simultaneous phenotypes using gTME. The tool of gTME was applied to the problem of imparting the multiple, simultaneous phenotype in *E. coli* of tolerance to both ethanol and SDS. Four distinct, alternative strategies were chosen to search for the best sigma factor mutant. A fitness assay was conducted whereas the best mutant was assayed for improved growth rate over the control in each of the nine conditions. The total fitness represents the cumulative sum of components in a matrix of fraction increase in growth-rate over control for these nine conditions. Component fitness (either ethanol or SDS) represent the summation of only conditions in which one of the component is varied, while the other is absent. It is possible for any of these fitness values to be negative when the mutant strain has a decreased growth rate compared with the control.

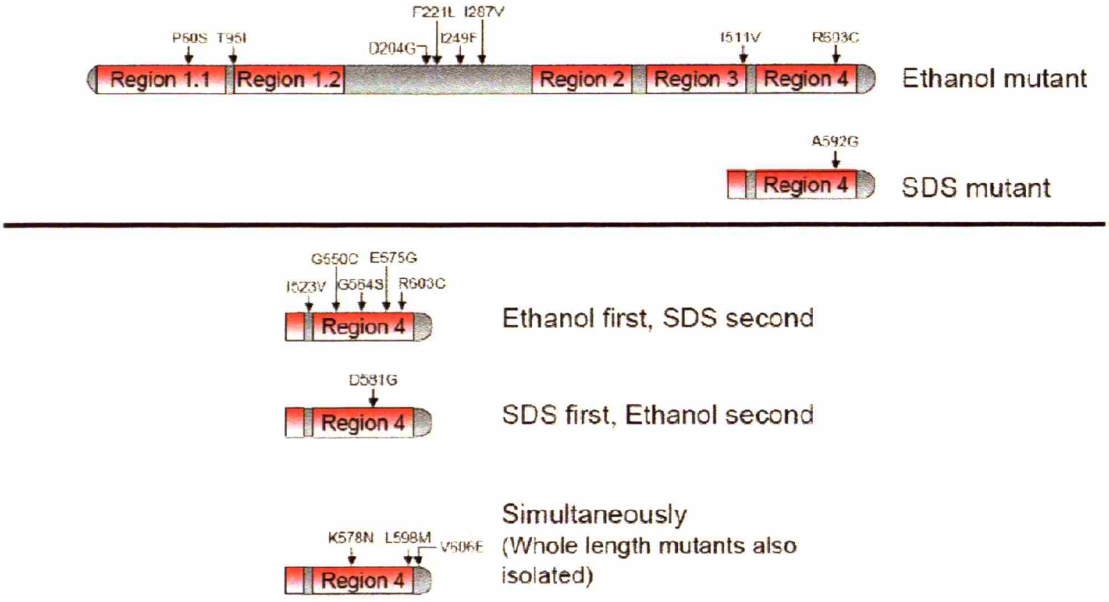

Figure 8.13: Sequence analysis of multiple, simultaneous phenotypes using gTME. Schematics for the best identified mutants identified for SDS and ethanol phenotypes using the four strategies illustrated in **Figure 8.12** are provided.

8.4.4 Acetate tolerance

Acetate is an *E. coli* byproduct that is inhibitory to cell growth and ultimately product formation in many important fermentations. High acetate levels in fermentations (especially above 10 g/L) are a common problem leading to an inhibition of cell growth and product formation (Lasko et al., 1997; Lasko, Zamboni, & Sauer, 2000). The mutant sigma factor library was serially subcultured twice on 20 g/L followed by 30 g/L of acetate in M9-minimal medium. Single colonies were isolated from this mixture, retransformed to preclude any chromosome-based growth adaptation, and assayed for growth in varying acetate concentrations. **Figure 8.14** compares the growth rates of the five re-transformed mutant strains with that of the control at 0, 10, 20, and 30 g/L of acetate. Isolated strains showed a drastic increase in tolerance (as measured by growth rate) in the presence of high levels (20 and 30 g/L) of acetate. At 30 g/L of acetate, the strains engineered through gTME had doubling times of 10.5 – 12.5 hours, approximately 1/5 of the doubling time of the severely inhibited control (56 hours doubling time). Furthermore, the growth rate of these improved mutants was not substantially affected in the absence of acetate (**Figure 8.14, 0 g/L**), which was a similar finding for previously identified strains with improved ethanol tolerance.

The underlying sequences of mutant sigma factors giving rise to acetate tolerance were analyzed. **Figure 8.15** summarizes the various mutations classified by region (Gruber & Gross, 2003) in the isolated sigma factors eliciting an increased cellular tolerance for acetate. Only one of the five isolated mutants was truncated. It was previously shown that truncated mutant sigma factors truncation arose from an artifact in

the restriction enzyme digestion and primer sequence similarities, and includes part of region 3 and the complete region 4 of the protein. Two residues are mutated in separate mutant sigma factors. First, the M567V mutation appeared in two of the acetate mutants (truncated, Ac-1 mutant and full-length Ac-3 mutant). Additionally, the I127 residue was mutated in two full-length versions, Ac-2 and Ac-4, however was changed to distinct amino acids, an asparagine and valine respectively. The bulk of the mutations appear to be distributed among the functional domains of the sigma factor. It is interesting to note that even though strains have similar tolerance profiles, the underlying mutations are diverse. These results suggest either different molecular mechanisms able to influence the same transcription profiles, or different transcriptional profiles responsible for improving acetate tolerance. Regardless of the mechanism, these strains present significant improvements in acetate tolerance over those previously reported in literature, with growth rates at the 20-30 g/L level comparable with other species of bacterium which are evolutionarily adapted for high acetate levels (Lasko, Zamboni, & Sauer, 2000).

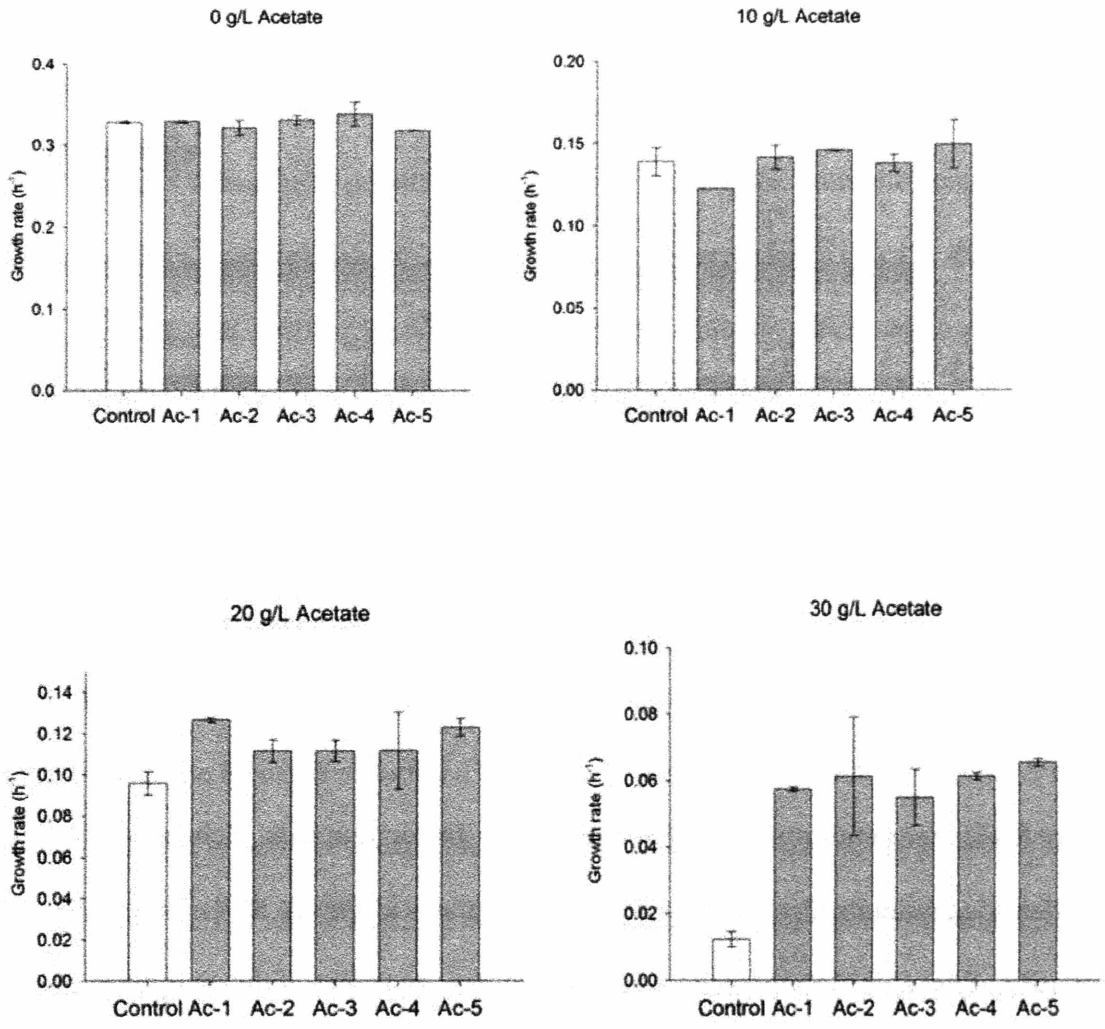

Figure 8.14: Growth analysis of acetate mutants. Strains harboring mutant sigma factors were isolated with increased tolerance to elevated levels of sodium acetate in minimal medium. The growth rate of the control strain and engineered strains were measured at 0, 10, 20, and 30 g/L of acetate respectively. While these mutants exhibit drastically improved growth rates at elevated levels of acetate, there is no reduction in the basal-level growth rate of the strain in the absence of acetate.

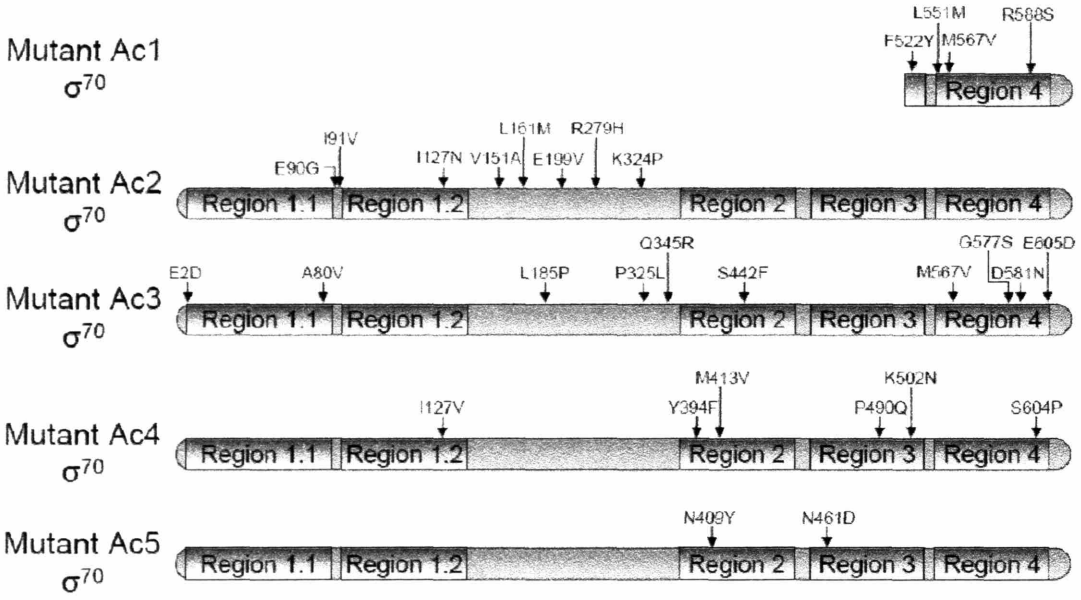

Figure 8.15: Sequence analysis of acetate sigma factor mutants. Schematics for the best identified mutants identified for the acetate mutants in **Figure 8.14**. Despite the similarity of the global phenotype (growth rates), the sequences of the underlying mutant sigma factors are quite diverse. The I127 and M567 residues are changed in multiple mutant sigma factors. Mutant Ac-1was a truncated sigma factor.

8.4.5 pHBA tolerance

Cellular tolerance to p-hydroxybenzoic acid and similar aromatic compounds is relatively low in *E. coli* which limits the prospect of bioproduction (Barker & Frost, 2001; Van Dyk et al., 2004). To identify sigma factor mutants which can improve tolerance to pHBA, the library was cultured in the presence of 20 g/L of pHBA overnight to select for strains with increased tolerance to this compound in terms of growth and viability at high pHBA concentrations. One strain was isolated with marked improvement in the growth yield at 13 hours compared with the control after re-transformation of the plasmid (**Figure 8.16**). The improvement in growth yield is more pronounced at the higher levels of pHBA. Growth yields after 15-20 g/L of pHBA are severely reduced in both the mutant and control strains. Once again, the growth yield in the absence of pHBA was unaffected compared with the control strain. Mutant HBA1 showed a truncated form of the sigma factor with a total of six mutations (**Figure 8.17**), with 4 of 6 residues being changed to a valine. Interestingly, no full-length mutants were isolated after the initial round of screening.

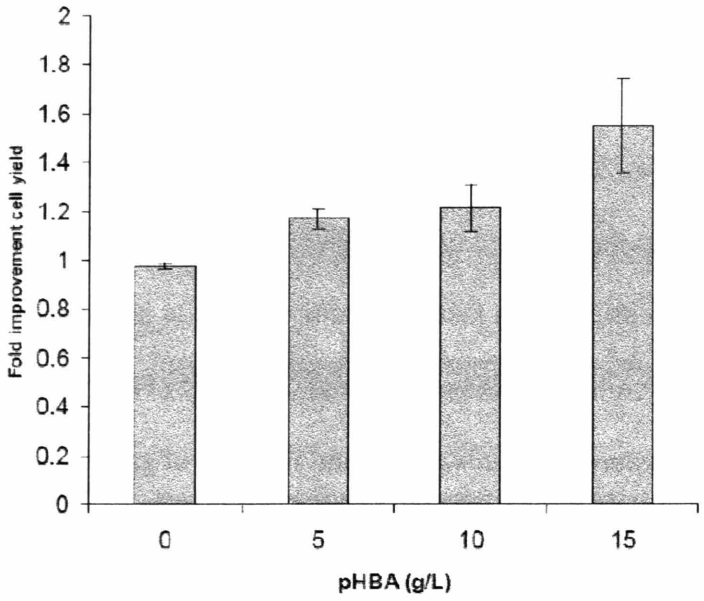

Figure 8.16: Growth analysis of pHBA sigma factor mutants. Strains harboring mutant sigma factors were isolated with increased tolerance to elevated levels of pHBA in minimal medium. The fold improvement in growth yield after 13 hours was measured between the mutant strain and the control strain at 0, 5, 10, and 15 g/L of pHBA. Normal growth was seen in the absence of pHBA.

Figure 8.17: Sequence analysis of pHBA sigma factor mutants. The sigma factor mutant, HBA-1, was a truncated form of the sigma factor possessing a total of 6 mutations, 4 of which are the conversion of amino acids into valines.

8.4.6 Hexane tolerance

As a final example of the capacity of gTME to improve bacterial phenotypes, n-hexane was used as a model for solvent tolerance. The phenotype of solvent tolerance is complex and as such, many genes have been identified which participate in tolerance (Abe et al., 2003; Aono, Negishi, & Nakajima, 1994; Shimizu et al., 2005). Furthermore, bacterial strain tolerance to organic solvents is useful for a variety of biotechnology applications, and as a result, has been explored using many of the traditional methods of cellular engineering. As a result, the tool of gTME was investigated for the potential to increase solvent tolerance in *E. coli*. The original *rpoD* (σ^{70}) mutant library was cultured and harvested in exponential phase and transferred to a two-phase system containing LB medium and hexane (10% v/v). Strains were isolated after 18 hours of growth in the presence of hexane. These individual colonies were again cultured to exponential phase and then cultured in the presence of hexane. Cell densities are measured after 17 hours. Two strains of interest showed enhanced growth yields in the presence of hexane after retransformation. Specifically, mutant Hex-1 showed a nearly 2.5-fold improvement in cell yield and mutant Hex-2 showed a nearly 2-fold improvement over the control (**Figure 8.18**). These two different mutants have very distinct sequences. In particular, Hex-1 was a full-length sigma factor with 5 mutations, 4 of which are in the non-conserved region. The Hex-2 mutant was truncated with only a single mutation at the Q589 residue. In each of these cases, the introduction of a mutant sigma factor resulted in the improvement of hexane solvent tolerance. Sequences are presented in **Figure 8.19**.

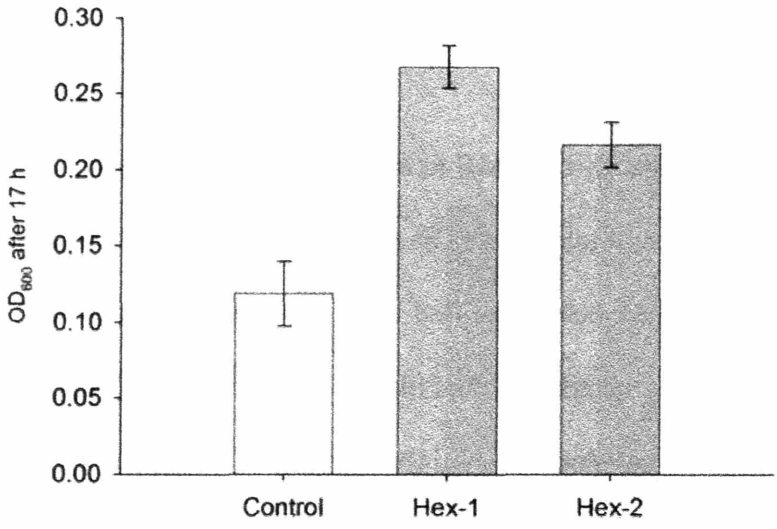

Figure 8.18: Growth analysis of hexane sigma factor mutants. Strains harboring mutant sigma factors were isolated with increased hexane in complex, LB medium. The growth yield (as measured by OD600) is shown for mutant and control strains after 17 hours of growth in a 10% v/v hexane saturated culture. The two mutants show a nearly 2.5 and 2 fold improvement in cell yield over the control.

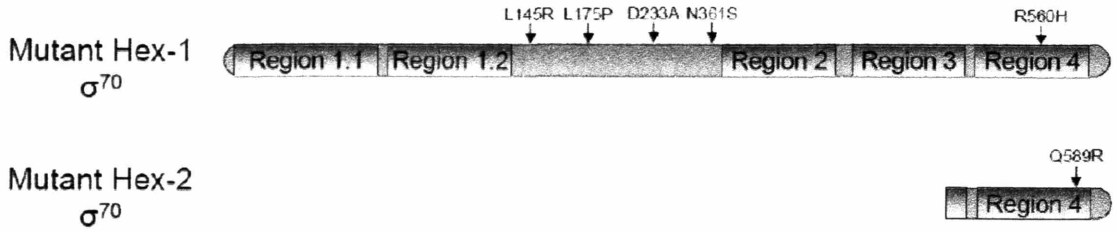

Figure 8.19: Sequence analysis of hexane sigma factor mutants. The sigma factor mutant Hex-1 (the mutant with the highest increase in growth yield), was a full-length sigma factor with 5 mutations, 4 of which are in the non-conserved region. The Hex-2 mutant was a truncated form possessing only a single amino acid substitution.

8.4.7 *E. coli* Summary

Collectively, and by virtue of their diversity and magnitude of achieved phenotype improvement, these *E. coli* examples illustrate the potential of gTME to mediate global transcriptome changes that allow organisms to access novel cellular phenotypes. In each of these cases, the improvements were significant and beyond the levels obtainable through traditional methods employed over decades of prior research. Yet, these examples represent only a small fraction of the broader potential of gTME. It is possible to futher explore this concept with other alternative sigma factors in *E. coli* and with a wide variety of additional phenotypes of interest. Furthermore, the method of gTME is not specific to prokaryotic cells, as the next section illustrates success in applying these concepts in yeast to obtain several phenotypes of biotechnological interest. In each of these examples, it is shown that the global changes brought about by random mutations in the components of transcriptional regulatory machinery improve cellular phenotypes beyond the levels attainable through rational engineering or traditional strain improvement by random mutagenesis.

8.5 *Yeast Applications*

In any type of cellular system, a subset of proteins is responsible for coordinating global gene expression. As such, these proteins provide access points for diverse transciptome modifications broadly impacting phenotypes of higher organisms. This section discusses the application of gTME to the eukaryotic model system of yeast

(*Saccharomyces cerevisiae*). In stark contrast to the transcriptional machinery of the prokaryotic system, eukaryotic transcription machinery is more complex in terms of the number of components and factors associated with regulating promoter specificity. First, there are three RNA polymerase enzymes with separate functions in eukaryotic systems while only one exists in prokaryotes. Furthermore, an example of this complexity is exemplified by nearly 75 components classified as a general transcription factor or coactivator of the RNA Pol II system (Hahn, 2004). Components of the general factor TFIID include the TATA binding protein (*Spt15*) and 14 other associated factors (TAFs) and are thought to be the main DNA binding proteins regulating promoter specificity (Hahn, 2004). Moreover, TATA-binding protein mutants have been shown to change the preference of the three polymerases, suggesting a pivotal role for orchestrating the overall transcription in yeast (Schultz, Reeder, & Hahn, 1992). The focus of this study will be on two major proteins of transcription: the TATA-binding protein (*Spt15*) and a TAF (*TAF25*).

Crystal structures are available for the TATA-binding protein and clearly illustrate portions of the protein for direct DNA binding and other portions for protein binding with the TAFs and parts of the polymerase (Bewley, Gronenborn, & Clore, 1998; Chasman et al., 1993; J. L. Kim, Nikolov, & Burley, 1993). This structure consists of two repeat regions which interact with the DNA and two helices which interact with proteins. Assays and mutational analysis suggest that the TATA-binding protein plays an important role in promoter specificity and global transcription. Furthermore, important residues have been suggested for DNA contact points and protein interaction points (Arndt et al., 1992; J. Kim & Iyer, 2004; Kou et al., 2003; Schultz, Reeder, & Hahn,

1992; Spencer & Arndt, 2002). The TAFs have received varying amounts of attention. The TAF25 protein, the subject of this study, has been analyzed using sequence alignment and through mutation analysis and has been shown to impact transcription of many genes (Kirchner et al., 2001). This protein is seen to have a series of helices and linkers which are critical to protein interactions. These proteins are investigated using the method of gTME to elicit three phenotypes of interest: (1) LiCl tolerance to model osmotic stress, (2) high glucose tolerance, and (3) the simultaneous tolerance to high ethanol and high glucose.

8.5.1 LiCl tolerance

Osmotic stress response and tolerance is a complex, pleiotropic response in cells. For yeast, it has been shown that elevated LiCl concentration can induce osmotic stress at concentrations around 100 mM (Haro, Garciadeblas, & Rodriguez-Navarro, 1991; J. H. Lee, Van Montagu, & Verbruggen, 1999; Park et al., 2003). Yeast cell libraries carrying the mutant versions of either the TBP or TAF25 were serially subcultured in the presence of 200 to 400 mM LiCl. Strains were isolated and retransformed to revalidate the phenotype was a result of the mutant factor. Interestingly, the best strains from each library showed varying improvements to LiCl. The TAF25 outperformed the respective TAF25 unmutated control at lower LiCl concentrations, but was not effective at concentrations above around 200 mM. Conversely, the SPT15 mutant was able to outperform the control at elevated levels of 150 to 400 mM. In each case, the growth phenotype in the absence of LiCl was not impacted by the presence of the mutant factor. A summary of the improvement in growth yield is provided in **Figure 8.20**. A sequence

analysis (**Figure 8.21**) indicates that the improvement in LiCl tolerance was controlled by a single mutation in each of the proteins, with the SPT15 mutation occurring in the unconserved region.

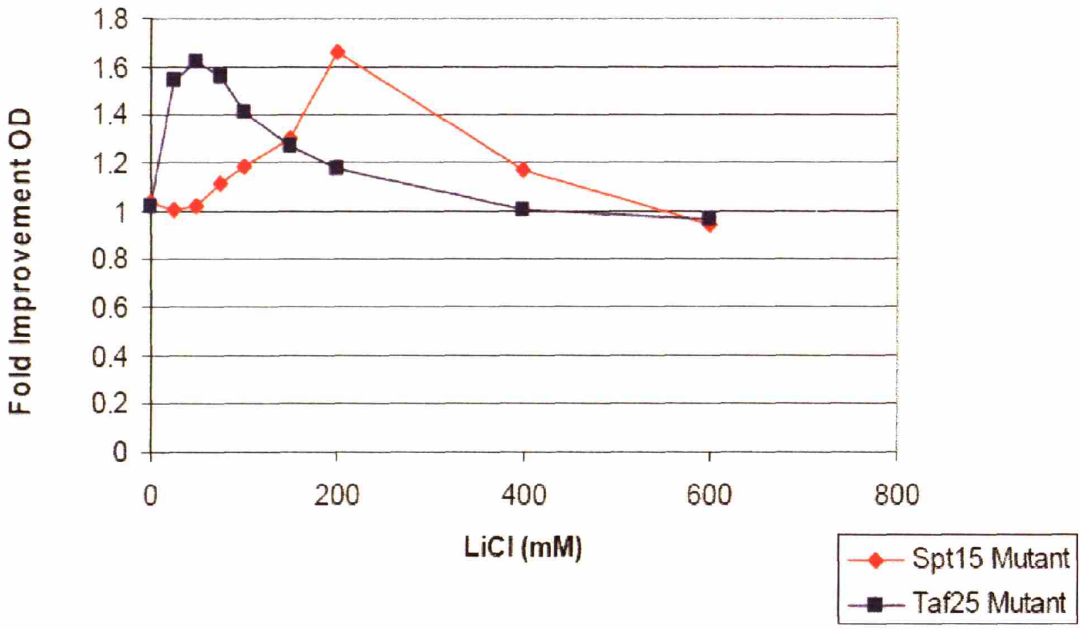

Figure 8.20: Growth analysis of LiCl gTME mutants in yeast. Strains harboring mutant *Taf25* or *Spt15* were isolated with through serial subculturing in elevated levels of LiCl in a synthetic minimal medium. The growth yield (as measured by OD600) is shown for mutant and control strains after 16 hours. The *Taf25* outperformed the control at lower concentrations of LiCl, while the *Spt15* mutant was more effective at higher concentrations.

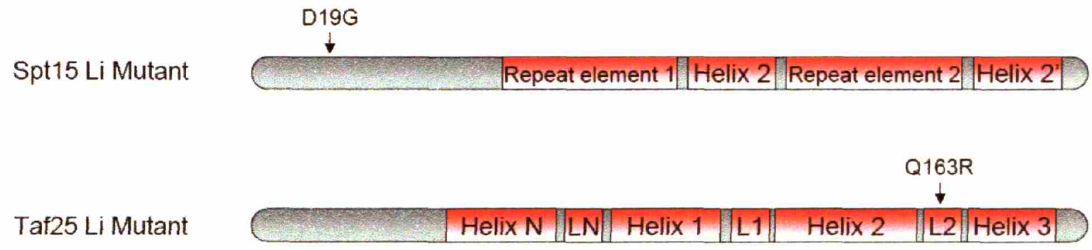

Figure 8.21: Sequence analysis of LiCl gTME mutants in yeast. Mutations are shown mapped onto a schematic showing critical functional components of the respective factor. Each mutant was seen to possess only a single amino acid substitution.

8.5.2 High glucose tolerance

High glucose fermentations have been explored for increasing the ethanol produced from a batch culture of yeast. However, these "very high gravity fermentations" are often quite inhibitory to cell growth and typically are treated by altering the medium composition, rather than altering the cells (Bafrncová et al., 1999; Bai et al., 2004; Thatipamala, Rohani, & Hill, 1992). To explore this problem using gTME, yeast cell libraries carrying the mutant versions of either the TBP or TAF25 were serially subcultured in the presence of 200 to 400 g/L of glucose. Strains were isolated and retransformed to revalidate the phenotype was a result of the mutant factor. Strains showed a 2 to 2.5 fold increase in cell density after 16 hours of culturing. Unlike the case with LiCl, both the TAF25 and SPT15 proteins showed a similar response to elevated glucose with the maximum improvement over the control occurring between 150 and 250

g/L. However, the SPT15 mutant showed a larger improvement over the TAF25. **Figure 8.22** presents the growth improvement of these mutants and the sequences are presented in **Figure 8.23**. In this case, both proteins had only a single mutation, however several suboptimal mutants were isolated for the SPT15 protein, some of which having as many as seven mutations. Both mutations shown here are located in known protein contact areas, especially the I143 residue in the TAF25 protein (Schultz, Reeder, & Hahn, 1992).

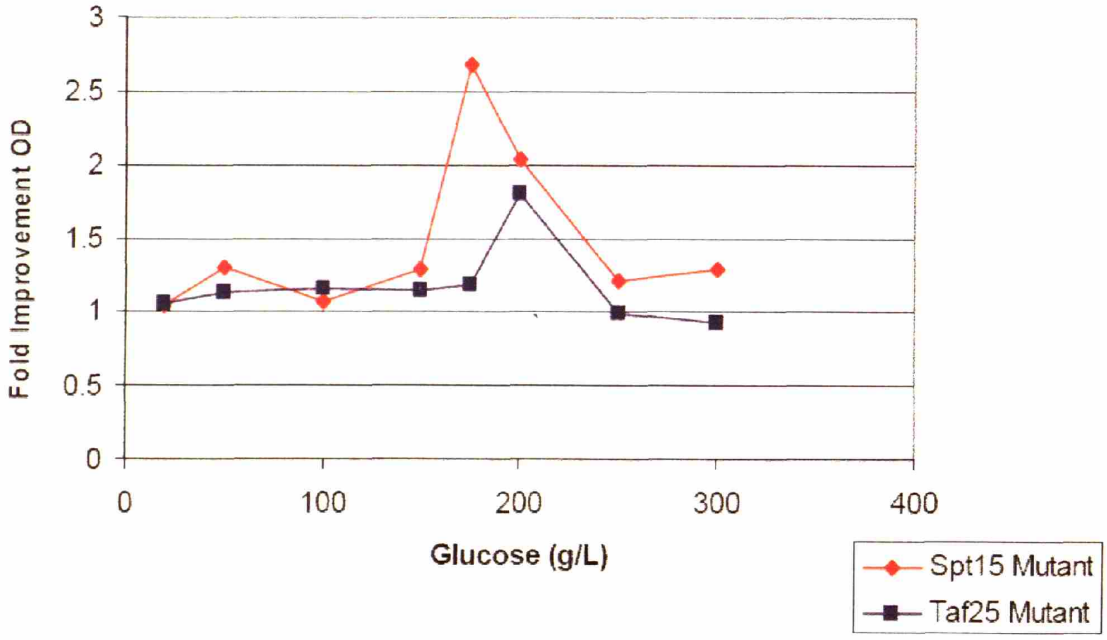

Figure 8.22: Growth analysis of glucose gTME mutants in yeast. Strains harboring mutant *Taf25* or *Spt15* were isolated through serial subculturing in elevated levels of glucose in a synthetic minimal medium. Here, both proteins show an improvement across a similar range of concentrations, with the SPT15 protein giving the largest improvement.

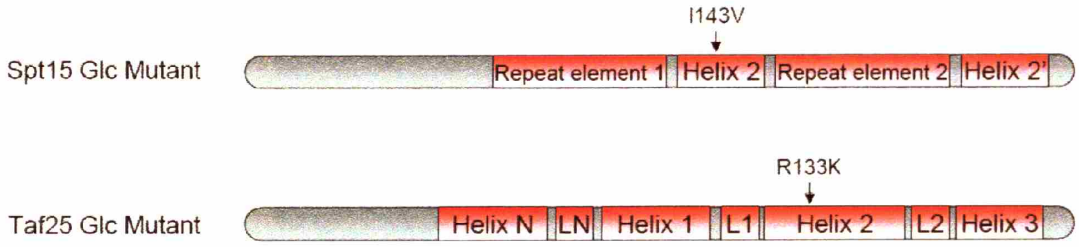

Figure 8.23: Sequence analysis of glucose gTME mutants in yeast. Mutations are shown mapped onto a schematic showing critical functional components of the respective factor. Each mutant was seen to possess only a single amino acid substitution, however several other SPT15 proteins were isolated, some possessing many mutations.

8.5.3 Ethanol and glucose multiple tolerance

Successful fermentations of bioethanol for yeast require tolerance to both high glucose and ethanol concentrations. To this end, the multiple tolerance phenotype was tested through the simultaneous treatment of both mutant libraries to elevated levels of ethanol and glucose (5% and 100 g/L). Isolated strains were retransformed and assayed under a range of glucose concentrations in the presence of 5 and 6% ethanol. Interestingly, the SPT15 mutants outperformed the control at all concentrations tested, upwards of 13 fold improvement in some concentrations. This improvement far exceeded the overall improvement of the TAF25 mutant which was not able to grow in the presence of 6% ethanol. **Figure 8.24** highlights the growth analysis of these best strains and sequences are provided in **Figure 8.25**.

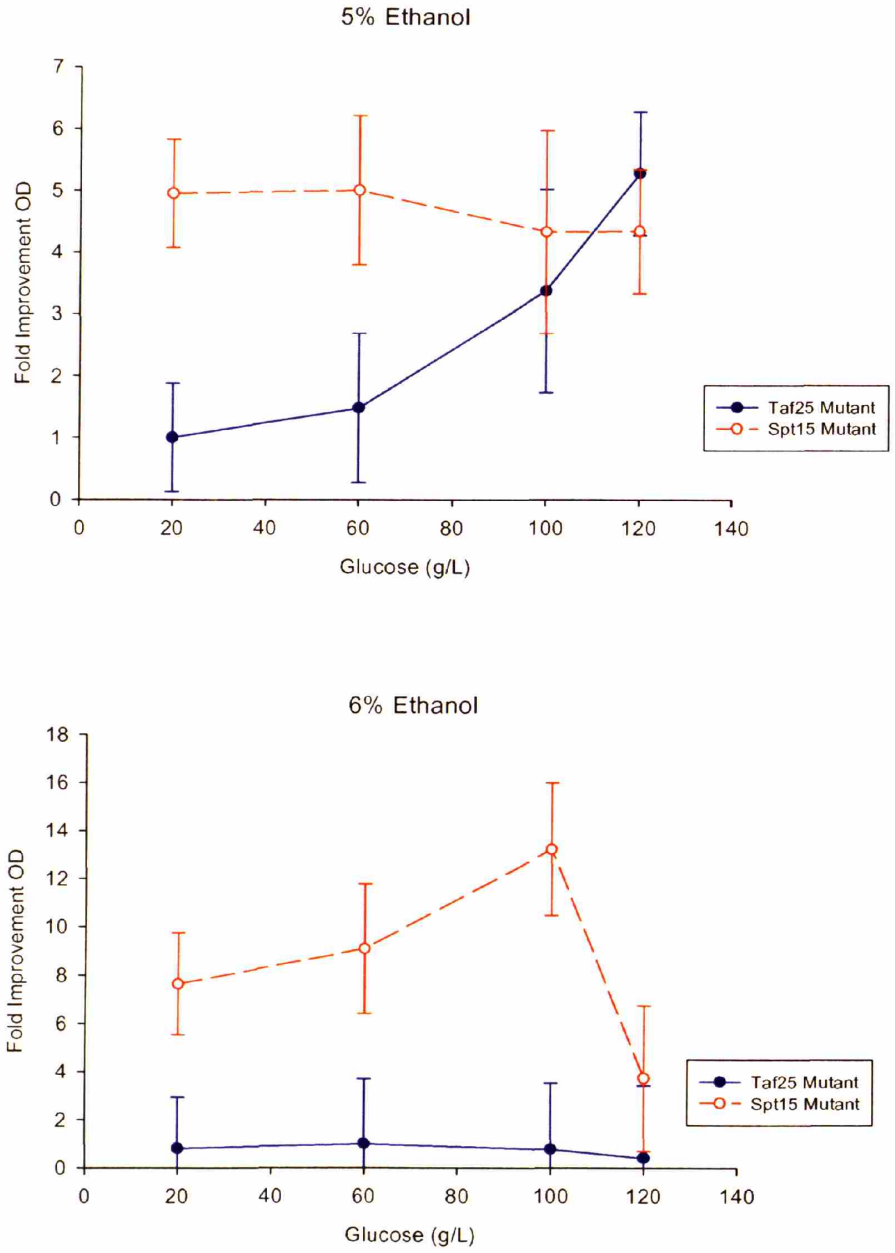

Figure 8.24: Growth analysis of ethanol-glucose gTME mutants in yeast. Strains harboring mutant *Taf25* or *Spt15* were isolated with through serial subculturing in elevated levels of ethanol and glucose in a synthetic minimal medium and assayed for growth at 20 hours. Here, the SPT15 protein far exceeded the impact of the TAF25 mutant.

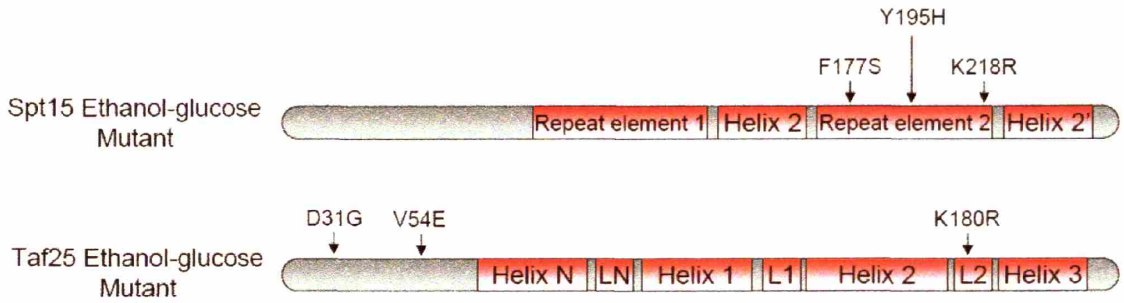

Figure 8.25: Sequence analysis of ethanol-glucose gTME mutants in yeast. Mutations are shown mapped onto a schematic showing critical functional components of the respective factor. Each mutant was seen to possess several single amino acid substitutions in critical regions for DNA or protein contacts.

8.6 *Summary*

The use of this tool of global Transcription Machinery Engineering (gTME) can unlock latent cellular potentials and create cells with superior phenotypic characteristics. Through the introduction of a mutant sigma factor into *Escherichia coli*, it was possible to elicit several important and complex phenotypes including the increase of ethanol tolerance. In particular, this tool was able to obtain strains capable of growing in upwards of 70 g/L of ethanol without negatively impacting the growth rate in the absence of ethanol, which exceeds the results of decades of research in this field. In a preliminary transcriptional analysis, it was found that ethanol tolerance is a highly pleiotropic phenotype controlled by a multitude of genes, illustrating the power of gTME to make global, simultaneous transcriptional level modifications. Each of the examples illustrated here have been explored in various capacities using a diverse array of tools for cellular

engineering. Typically, these tools must be performed in a sequential manner with smaller, discretized improvements in phenotype at each stage. Through the use of gTME, it was possible to identify mutants possessing significantly enhanced phenotypic characteristics.

Despite complexity of eukaryotic systems, it was demonstrated that modifications of the subsets of proteins responsible for controlling transcription were effective in eliciting powerful phenotype changes. Since control is more distributed in a eukaryotic system, it may be possible to achieve further improvements by creating libraries encompassing all of the possible proteins involved in the transcription machinery. Regardless, it was possible to make changes to cellular phenotype through the introduction of a few mutations into critical proteins associated with global promoter recognition. More importantly, these results highlight the generic nature of the gTME approach and serve as a proof-of-concept for other eukaryotic and higher-level ystems.

Furthermore, each of the phenotypes obtained were regulated by a single, mutant protein which may be subjected to an additional round of mutation and selection. It was shown through an example with ethanol tolerance that subsequent rounds of mutagenesis and selection can have a profound effect of significantly enhancing the phenotype of interest. These results suggest that the phenotypes achieved in this study may be further optimized. Additionally, the introduction of full and truncated versions of mutant sigma factors may provide a means of further increasing tolerance. This co-expression was found to be optimal for providing cells with a dual phenotype of ethanol and SDS tolerance. Finally, these strains may be further analyzed on the basis of transcriptional profiling to gain an understanding of mechanism and the number of genes responsible for

regulating these cellular tolerances. Regardless of the follow-up and analysis, these results highlight the potential use of gTME to improve cellular phenotype and to identify mutant strains with superior properties for use in diverse bioprocessing applications.

For the first time, these results demonstrated the application of global Transcription Machinery Engineering to alter cellular phenotype. As such, the gTME paradigm allows for cellular and metabolic engineering to be reduced to a problem of protein evolution. This strategy allowed for the directed modification of the genetic control of multiple genes simultaneously, as opposed to typical consecutive, gene-by-gene strategies. Furthermore, we found that multiple rounds of gTME allowed sequential phenotypic improvements by probing deeper into the vast sequence space of transcription factor engineering. As a result, it is now possible to unlock complex phenotypes regulated by multiple genes which are essentially unreachable by the relatively inefficient iterative search strategies. It is worth noting that the described method can also be applied in reverse to uncover complicated genotype-phenotype interactions, as illustrated by the results of the ethanol tolerance study. In such applications, one would employ a number of high-throughput cellular and molecular assays to assess the altered cellular state and ultimately deduce systematic mechanisms of action underlying the observed phenotype in these mutants. We also envision that this tool can be used to uncover mechanisms responsible for imparted, complex phenotypes such as disease states. Hence, gTME as described here is a paradigm shifting method for identifying genetic targets, eliciting desired phenotypes, and realizing the goal of whole cell engineering.

Chapter 9

Conclusions and recommendations

9.1 Summary

Systematic and combinatorial methods were investigated for the identification of genetic targets for improved phenotypes including the production of value-added products and increasing cellular tolerances. Initially, a global stoichiometric model was used to identify important single and multiple gene knockout targets for lycopene production in *Escherichia coli* (Chapter 4). Furthermore, these models enabled the investigation of many putative parameters impacting lycopene production including oxygen uptake rates and byproduct formation (Section 4.2). These targets led to substantial increases in lycopene production, but are limited due to the nature of these models, which lack information about regulation and kinetics. Despite these limitations, many of the key genetic nodes found through the generation of the gene knockout search network (Chapter 6) coincided with the targets found from the stoichiometric model.

To perform a more complete search and further increase production levels, these systematic approaches and targets were complemented with combinatorial searches to

identify unknown and regulatory targets aided by global transposon mutagenesis. When combined, these searches led to further increases of lycopene production exceeding levels previously reported in literature and allowed, for the first time, the visualization of the metabolic landscape formed by combining these two disjoint set of genetic targets (Section 5.2). This metabolic landscape was nonlinear and complex. Further analysis of the landscape suggested that sequential searches for genetic targets may be limited to metabolic or localized targets and not generally transferable across genotypes, as the impact of combinatorial targets (often regulatory factors) are introducing nonlinearity (Section 5.3). Culturing of the two global maxima strains and the parental strain in high cell density fermentations led to the quantification of differences between the two engineered strains and resulted in the accumulation of 220 mg/L of lycopene in 24 hours (Section 5.5).

These investigations and subsequent analysis aided in understanding the sources of non-linearity in the metabolic landscape and provided broad strategies for dealing with two distinct sets of gene knockouts (stoichiometric and regulatory/unknown). However, it is unknown how the overall search trajectory biases the exploration of the metabolic landscape, and ultimately confounds the search for global maxima. To address this issue, transposon mutagenesis was used in the background of eight different genotypes spanning interesting strains in the metabolic landscape (Section 6.1). These searches led to the formation of the gene knockout search network (Section 6.2). These results illustrated that while metabolic networks are complex, a small subset of key "gateway" genetic targets helps unlock the cellular phenotype and provide invaluable leads for phenotype optimization (Section 6.3). Furthermore, in depth characterization of targets

such as *yliE* indicates that genetic targets are specific to both environmental/culturing conditions and genotype (Section 6.4.1). Finally, as a result of the topology of the landscape, a simple optimization algorithm such as the greedy algorithm was unable to solve the phenotype optimization problem and would have led to a suboptimal solution.

To further address the issues of target identification and controlled expression, the tools of promoter engineering (Chapter 7) and global Transcription Machinery Engineering (gTME) (Chapter 8) were developed. These two examples illustrate that the directed mutagenesis of carefully chosen, functional components of DNA or proteins can alter function and phenotype drastically. To create a tool for the controlled expression of genes, a heterologous promoter was subjected to error-prone PCR and subsequently screened and characterized for promoter function (Section 7.2 and 7.3). These promoters were integrated into the genome to allow for a quantitative characterization of genotype-phenotype relationships. The examples of growth yield as a function of *ppc* level and lycopene production as a function of *dxs* level highlight that gene expression is a continuum which can exhibit well defined, gene and genotype-specific maxima (Section 7.4). Furthermore, this tool may be useful for the introduction of gene knockdowns as seen from the example with *ppc* (Section 7.4.3). Collectively, a library of well-defined, homogeneous promoters allows for the efficient tuning of genetic control.

Metabolic and cellular landscapes are too complex and nonlinear to efficiently and completely probe through single gene modifications linked with a search strategy, even with tools for controlled gene expression such as promoter engineering. Furthermore, many important phenotypes are regulated by multiple genes and it was shown that the engineering of global transcription machinery could create the necessary

simultaneous modification of genes (Chapter 8). Through the introduction of engineering transcription machinery, it was possible to elicit complex tolerance and metabolic phenotypes in both *E. coli* (Section 8.4) and *S. cerevisiae* (Section 8.5) which highlights the generic nature of this approach. Furthermore, transcriptional profiling may allow for this tool to be used in reverse to allow for the extraction of the underlying molecular mechanisms responsible for the elicited phenotype. Finally, this approach relieves the limitation of single gene searches and may be iterated since the phenotype is being regulated by a single mutant protein, which may be subjected to subsequent rounds of mutagenesis.

Collectively, these results provided evidence for the utility and status of systematic and combinatorial approaches for the metabolic engineering of microorganisms. Furthermore, many of the strategies and tools may be applied to other cellular systems of interest beyond microorganisms.

9.2 Conclusions

Explorations and optimizations through the genomic space are a daunting undertaking given the complexity and size of the possible search space. The results discussed in this thesis highlight that these spaces are highly non-linear exhibiting multiple optima. These searches are similar in size as protein directed evolution problems; however, they suffer from less developed tools and more complex component interactions. In the latter case, mutations which increase protein function are typically

additive in nature (Wells, 1990; Zhang et al., 1995). On the other hand, the mutations and genotypes incurred in altering global cellular phenotypes are not necessarily additive and can be quite non-linear. In this regard, many local maxima may occur in a phenotype space due to the various subsets of gene alterations which may lead to improved phenotypes.

Metabolic landscapes are not unlike solving a complex optimization problem. Nonlinear, global optimization is a difficult problem to solve computationally. Many commonly used, simple techniques such as methods of steepest accent are often not adequate to guarantee convergence to the global maxima and can be quite sensitive to initial conditions. Optimizations within the metabolic landscape suffer from similar problems. Following a greedy-algorithm for strain optimization may work when targets are additive, as exhibited with the stoichiometric modeling, but will fail in the nonlinear landscape carved by including regulatory and unknown factors. Various computational models and approaches such as simulated annealing, branch-and-bound, and perturbation theorems, especially randomized initial conditions, have been successful in handling these difficult mathematical spaces. In the case of optimizing cellular phenotype, combinatorial tools provide for the ability to probe various combinations and serve to introduce perturbations of starting points. Short of exhaustive searches, which become infeasible when dealing with multiple modifications, combinatorial tools linked with a high-throughput selection provide a more efficient means of creating the broad-ranging modifications necessary to obtain global maxima. Nonetheless, systematic tools are still quite useful for identifying key initial genotypes for combinatorial exploration.

The development of tools for controlled gene expression (aided through promoter engineering) and multiple gene modification (aided through global Transcription Machinery Engineering) greatly assist the effort of engineering cellular function. Collectively, these tools allow for novel and directed manipulations and modifications which aid in the exploration of these landscapes. Despite the power and utility of these tools, their effectiveness depends on the search strategy invoked to optimize phenotype. As demonstrated through the results presented in this thesis, the concepts of target identification, target modification, and search strategy lie at the heart of metabolic and cellular engineering endeavors and are critical components for unlocking cellular potential.

In summary, the following set of ten major conclusions may be drawn from this work:

- Collectively, the tools and approaches demonstrated here can help realize the goal of whole-cell engineering.
- While metabolic networks are complex, a small subset of key "gateway" genetic targets helps unlock the cellular phenotype.
- Stoichiometric models are helpful in guiding the search of the metabolic landscape and identifying putative parameters leading to a given phenotype, however are limited in ability to extract all targets.
- Sequential searches for genetic targets may be limited to metabolic or localized targets and not generally transferable across genotypes.

- Search algorithms relying solely on a greedy algorithm may fail to optimize cellular or metabolic phenotypes.
- Gene expression is a continuum which can exhibit, well-defined genotype-phenotype relationships.
- Optimality is gene-specific and thus well-defined and characterized tools for genetic control are essential for quantifying the genotype-phenotype relationship.
- Metabolic and cellular landscapes are too complex and nonlinear to efficiently and completely probe through single gene modifications linked with a search strategy and thus require tools for engineering cells at the global level.
- Complex, important phenotypes are regulated by multiple genes and the engineering of transcription machinery can elicit the multiple, simultaneous modifications necessary to access these cellular phenotypes.
- Combinatorial methods are generally more powerful in obtaining a given cellular objective than systematic methods due to their ability to make broader perturbations. However, properly designed search strategies which make use of both systematic and combinatorial approaches may be the best route for optimizing phenotypes.

9.3 Recommendations for future work

There are several areas of future work which can extend from the results described in this thesis. Many of the phenotype-genotype investigations focused on the model system of lycopene production in *E. coli*. By natural extension, it would be

intriguing to perform similar studies for diverse organisms for divergent products. By extending to different products, it would be possible to assess the strength of the various tools for a given pathway. In the example used here, lycopene accumulates in the membranes of cells, is not metabolized, and does not account for a large fraction of the carbon balance. It is unclear how well a stoichiometric model will perform for a large biopolymer which accounts for a large fraction of the carbon balance. Furthermore, other cellular systems such as yeast and fungal systems are important organisms in the biotechnology sector and are unexplored with respect to metabolic landscapes. Additional understanding and examples are necessary before creating generic, broad-ranging rules for efficient search strategy techniques for strain improvement.

The results of gene knockout search network highlighted that many, varied genotypes could yield the same phenotype. However, what is still unknown is whether these strains, which appear different based on genotype, are acting through the same molecular mechanism. To this end, transcriptional data may be coupled with pathway analysis to elucidate the key genetic targets responsible for the phenotype of interest.

Finally, the results and extension of gTME into eukaryotic systems, highlighted by the successes in yeast illustrate that the technique is generic in nature. Therefore, further eukaryotic systems as well as extension in prokaryotic systems may be studied in more depth. In particular, the size and breadth of the gTME library can be extended and enhanced through applying gene shuffling between heterologous hosts, including all possible truncated forms of proteins, and creating synthetic chimeras of domains from different proteins. Furthermore, vectors can be constructed which co-express two or more mutant factors simultaneously. In general, the tools and strategies described in this

work have been developed to a level of utility. Now, it is necessary to refine and test these tools for varied systems.

Chapter 10

Materials and methods

10.1 Commonly used techniques

10.1.1 Flux balance analysis calculations

Application of FBA to the carotenoid system required including the non-endogenous reactions (*crtEBI*) required for the production of these molecules on the background of previously published stoichiometric models (Edwards & Palsson, 2000; Segre, Vitkup, & Church, 2002) with alterations in the isoprenoid biosynthesis pathway (see supplemental information). The resulting model consisted of 965 fluxes involving 546 metabolite intermediates. This model was solved subject to MOMA using the linear and quadratic programming methods using a PERL script (Edwards & Palsson, 2000; Segre, Vitkup, & Church, 2002). An additional script was used to perform the genome-wide knockout searches. For the calculation parameters, values for the glucose uptake, oxygen uptake and nitrogen uptake were set at 5, 200, and 1000 respectively. These values allow for glucose to be the limiting substrate in these calculations. Single knockout calculations

were performed on a Pentium IV Linux platform while the exhaustive double knockout search was performed on six Power PC 1.5 GHz microchips on an AIX platform.

10.1.1 Lycopene Assay

Intracellular lycopene content was extracted from 1 ml of bacterial culture at the point of total glucose exhaustion. The cell pellet was washed, and then extracted in 1 ml of acetone at 55°C for 15 minutes with intermittent vortexing. The lycopene content in the supernatant was quantified through absorbance at 475 nm (Seon-Won Kim & Keasling, 2001) and concentrations were calculated through a standard curve. The entire extraction process was performed in reduced light conditions to prevent photo-bleaching and degradation. Cell mass was calculated by correlating dry cell with OD600 for use in ppm calculations.

10.1.1 Transposon library generation and screening

Transposon libraries were generated using the pJA1 vector (Badarinarayana et al., 2001). Cells were transformed with between 800 and 1600 ng of the plasmid, then plated on the appropriate medium (LB or minimal) supplemented with 20 μM IPTG and the appropriate concentration of antibiotics. Plates were incubated at 37 °C for 16 - 36 hours as needed for colony development, and then allowed to sit at room temperature. Cells identified as exhibiting increased lycopene content (more red) were isolated and cultured. The identity of promising targets were sequenced using an abbreviated version of

Thermal Asymmetric Interlaced PCR (TAIL-PCR) (Liu & Whittier, 1995). For the TAIL1 reaction, 1.5 uL of genomic DNA isolated using the DNA purification kit (Promega) was used as the initial template. The TAIL3 reaction was increased to 30 cycles. Kanamycin specific primers: TAIL1–TATCAGGACATAGCGTTGGCTACCCG, TAIL2–CGGCGAATGGGCTGACCGCT, TAIL3 – TCGTGCTTTACGGTATCGCCGCTC. The degenerate primer AD1 was used as described in the reference. The product of the TAIL3 reaction was purified by a PCR cleanup kit (Qiagen) after gel visualization. This product was sequenced using the primer TAIL-seq–CATCGCCTTCTATCGCCTTCTT.

10.1.1 Gene knockout construction and verification

Gene deletions were conducted using PCR product recombination (Datsenko & Wanner, 2000) using the pKD46 plasmid expressing the lambda red recombination system and pKD13 as the template for PCR. Gene knockouts were verified through colony PCR. Phage transduction was used for creating multiple gene knockout strains. P1*vir* phage transduction was used to transfer knockout mutants between strains (Miller, 1992).

10.2 Systematic gene knockouts

10.2.1 Strains and media

E. coli K12 PT5-dxs, PT5-idi, PT5-ispFD, provided by the DuPont Company, was used as the lycopene expression strain when harboring the pAC-LYC (Cunningham FX Jr, 1994) plasmid containing the *crtEBI* operon. Gene deletions were conducted using PCR product recombination (Datsenko & Wanner, 2000) using the pKD46 plasmid expressing the lambda red recombination system and pKD13 as the template for PCR (see supplemental information for primer designs). Gene knockouts were verified through colony PCR. Strains were grown at 37°C with 225 RPM orbital shaking in M9-minimal media (Maniatis, 1982) containing 5 g/L d-glucose and 68 μg/ml chloramphenicol. All cultures were 50 ml grown in a 250 ml flask with a 1% inoculation from an overnight 5 ml culture grown to stationary phase. All experiments were performed in replicate to validate data and calculate statistical parameters. Glucose monitoring was conducted periodically using a YSI2300 glucose analyzer to verify complete usage of glucose. Cell density was monitored spectrophotometrically at 600 nm. All PCR products were purchased from Invitrogen and utilized Taq polymerase. M9 Minimal salts were purchased from US Biologics and all remaining chemicals were from Sigma-Aldrich.

10.2.2 Primers for gene knockouts

All gene knockouts were constructed through PCR product inactivation with pKD13 as the Kan template for PCR. The following sets of primers were used in the construction of PCR products to inactivate the respective genes. To verify recombination, internal primers (k1, k2 and kt) as described in the protocol reference were used along

with the listed external antisense verification primer. A list of primers is provided in **Table 10.1**.

Gene	Strand	Primer 5' – 3'
gdhA	Sense	AACCATGTCCAAAAGCGCGACCCGAATCAAACCGAGTTC GGTGTAGGCTGGAGCTGCTTC
	Antisense	TCACACCCTGCGCCAGCATCGCATCGGCAACCTTCACAAG GGATCCGTCGACCTGCAGTT
	Verification	GATAAGCGTAGCGCCATCAG
gpmA	Sense	TTATCAAGATATTTCACCAGCGCACGTAAAGAGTTACCGT GTGCAGCGATGATCATCCGTCGACCTGCAGTTCGA
	Antisense	TACGACGTGGATCTGTCTGAGAAAGGCGTAAGCGAAGCA AAAGCAGCAGGTAAGCGTGTAGGCTGGAGCTGCTTC
	Verification	TCGCATCAGGCAATGTGCTCCAT
gpmB	Sense	GGGCCAGTCTGACAGCCCGCTGACCGCCAAAGGTGAGCA A GTGTAGGCTGGAGCTGCTTC
	Antisense	CGGCGCTCTGCCCATGCTGGTAATCCGAGAATCGTACTCA TCCGTCGACCTGCAGTTCGA
	Verification	CCCAATTAATCTACGCTGTG
aceE	Sense	TAAATTCCTGAAATATCTGGAACACCGTGGCCTGAAAGA T GTGTAGGCTGGAGCTGCTTC
	Antisense	TGGAAGCCGAACATCGAGTAATAGATGTAGAACGGGATC A TCCGTCGACCTGCAGTTCGA
	Verification	ACGCTCCAGACCGTCATGCA
ppc	Sense	AGCTCAATACCCGCTTTTTCGCAGGTTTTGATTAATGCAT GTGTAGGCTGGAGCTGCTTC
	Antisense	GGAAGTGAACGCCTGTTTAAAACAGCTCGATAACAAAGA TGGATCCGTCGACCTGCAGTT
	Verification	GCAAAGTGCTGGGAGAAACCATCAA
talB	Sense	ATGACGGACAAATTGACCTCCCTTCGTCAGTACACCACCG GTGTAGGCTGGAGCTGCTTC
	Antisense	ACGGATACCTTCCGCCAGTTTATCTACTGCCATTGGATCC TCCGTCGACCTGCAGTTCGA
	Verification	TGATACACTGCGAAGGGAGTGACAGACAGG
fdhF	Sense	ACGGTAAAATAACATCCGCCGCCGACGCGGTTTTGGTCAT GTGTAGGCTGGAGCTGCTTC
	Antisense	GCATCAGGTTGCAAAATCAACCTGGTCGTCGATAACGGC A TCCGTCGACCTGCAGTTCGA
	Verification	CGCGGTATTTCGTTTCGTCA

Table 10.1: Primer Designs for Gene Knockout Constructs.

10.3 Metabolic landscape

10.3.1 Strains and media

E. coli K12 PT5-*dxs*, PT5-*idi*, PT5-*ispFD*, provided by DuPont, was used as the lycopene expression strain when harboring the pAC-LYC plasmid containing the *crtEBI* operon (Cunningham FX Jr, 1994). Over-expressions of *dxs*, *idi*, and *ispFD* were chromosomally incorporated without an antibiotic marker through promoter delivery. Strains were grown at 37°C with 225 RPM orbital shaking in M9-minimal media (Maniatis, 1982) containing 5 g/L D-glucose and 68 μg/ml chloramphenicol. All simple cultures were 50 ml, grown in a 250 ml flask with an 1% inoculation from an overnight 5 ml culture and assayed at 15, 24, 39, and 48 hours. Optimized shake-flasks were 50 ml cultures grown in 250 ml flasks with a 1% inoculation from an overnight 5 ml culture with glucose feeds of 5 g/L at 0 and 15 hours and 3 g/L at 24 hours. The media for these experiments were M9-minimal media (Maniatis, 1982) with double concentrations of all salts except $CaCl_2$ and $MgSO_4$. All experiments were performed in replicate to validate data and calculate statistical parameters. Glucose monitoring was conducted periodically using r-Biopharm kit to verify complete usage of glucose. Cell density was monitored spectrophotometrically at 600 nm. All PCR products were purchased from Invitrogen and utilized Taq polymerase. M9 Minimal salts were purchased from US Biological and all remaining chemicals were from Sigma-Aldrich.

10.3.2 Hierarchical Clustering Routines

Hierarchical cluster was performed using complete linkage hierarchical clustering with a Euclidean distance similarity metric using Cluster Version 3.0. Dendrograms were visualized using Java TreeView Version 1.0.8.

10.4 High cell density fermentations

10.4.1 Fermentation conditions

All fed-batch fermentations were conducted in 1.5 L Applikon vessels containing an initial volume of 500 ml (for M9-based cultures) or 600 ml R-medium (Riesenberg et al., 1991) containing 5 g/L D-glucose and 68 μg/ml chloramphenicol.. A starting inoculum of 4% by volume (for M9-cultures) and 1% by volume (for R-media) was used. pH was measured and controlled online using NH_4OH and HCl, as appropriate for the desired control strategy. Temperature was regulated and controlled through water bath circulation. Glucose concentration was monitored and controlled online at a setpoint of 0.45 g/L using a YSI 2700 biochemistry analyzer connected to a pump with a feedstock of 200 g/L glucose with 0.1% antifoam (the set sampling interval was 20 minutes). Agitation speed was increased every two hours for the high cell density fermentations to obtain a linear increase from 400 RPM to 1100 RPM. Samples were taken every two hours for organic acid, amino acid, lycopene and biomass measurements. Online off-gas

analysis was performed using an MGA1600 mass spectrometer. Inlet air was supplied for aeration at a pressure of 15 psig.

10.4.2 Organic and amino acid measurements

A Bio-Rad Aminex HPX-87H reverse phase column with a Hewlett Packard 1050 HPLC system was used to measure acetate and formate. The column was run in an isocratic mode at a flow rate of 0.6 ml/min with a mobile phase consisting of 0.005 N H2SO4 at 45 °C. Detection was done by UV absorbance at 210 nm. Amino acids were analyzed as ortho-phthaldialdehyde(OPA) derivatives using an AminoQuant (Agilent Technologies, Palo Alto, CA) 2.1mm bore reversed phase column with a Hewlett Packard 1050 HPLC system which allowed full automation of derivatization, chromatography, data acquisition and data evaluation. Detection was done by UV absorbance at 338nm and the column was run at 40 °C in a gradient mode at the flow rate of 0.45 ml/min consisting of two buffers: BufferA contained 20 mM Na-Acetate, 50 mM tetrahydrofuran (HPLC-grade, Fluka Chemical Corp., Ronkonkoma, NY) and 2mM triethylamine (HPLC-grade, Fulka Chemical Corp.) at pH 7.2. Buffer B contained 20 mM Na-acetate pH7.2, methanol (HPLC-grade, Mallincroft Corp.), and acetonitrile HPLC-grade, Mallincroft Corp.) at a volumetric ratio of 20:20:40.

10.5 Probing the metabolic landscape

10.5.1 Strains and media

E. coli K12 PT5-*dxs*, PT5-*idi*, PT5-*ispFD*, provided by DuPont, was used as the lycopene expression strain when harboring the pAC-LYC plasmid containing the *crtEBI* operon (Cunningham FX Jr, 1994). Over-expressions of *dxs*, *idi*, and *ispFD* were chromosomally incorporated without an antibiotic marker through promoter delivery. Strains were grown at 37°C with 225 RPM orbital shaking in M9-minimal media (Maniatis, 1982) containing 5 g/L D-glucose and 68 μg/ml chloramphenicol. All simple cultures were 50 ml, grown in a 250 ml flask with an 1% inoculation from an overnight 5 ml culture and assayed at 15 and 24. Selected strains were grown in 50 ml cultures in 250 ml flasks with a 1% inoculation from an overnight 5 ml culture with glucose feeds of 5 g/L at 0 and 15 hours and 3 g/L at 24 hours. The media for these experiments were M9-minimal media (Maniatis, 1982) with double concentrations of all salts except $CaCl_2$ and $MgSO_4$. All experiments were performed in replicate to validate data and calculate statistical parameters. Glucose monitoring was conducted periodically using r-Biopharm kit to verify complete usage of glucose. Cell density was monitored spectrophotometrically at 600 nm. All PCR products were purchased from Invitrogen and utilized Taq polymerase. M9 Minimal salts were purchased from US Biological and all remaining chemicals were from Sigma-Aldrich. Transposon knockouts were generated as described in Section 10.1.1.

10.6 Promoter Engineering

10.6.1 Strains and media

E. coli DH5α (Invitrogen) was used for routine transformations as described in the protocol. E. *coli* K12 (MG1655) and *E. coli* K12 PT5-*dxs*, PT5-*idi*, PT5-*ispFD* (provided by DuPont) were used for promoter engineering examples. In specified strains the lycopene expression was performed using the pAC-LYC plasmid (Cunningham FX Jr, 1994) and assayed as described previously (Alper et al., 2005b). Assay strains were grown at 37°C with 225 RPM orbital shaking in M9-minimal media (Maniatis, 1982) containing 5 g/L D-glucose. When necessary, the M9 media was supplemented with 0.1% casamino acids. All other strains and propagations were cultured at 37°C in LB media. Media was supplemented with 68 μg/ml chloramphenicol, 20 μg/ml kanamycin, and 100 μg/ml ampicillin as necessary. Glucose monitoring was conducted using r-Biopharm kit. Cell density was monitored spectrophotometrically at 600 nm. All PCR products and restriction enzymes were purchased from New England Biolabs and utilized Taq polymerase. M9 Minimal salts were purchased from US Biological and all remaining chemicals were from Sigma-Aldrich.

10.6.2 Library construction

Nucleotide analogue mutagenesis was carried out in the presence of 20 μM 8-oxo-2'-deoxyguanosine (8-oxo-dGTP) and 6-(2-deoxy-β-D-ribofuranosyl)-3,4-dihydro-8H-pyrimido-[4,5-c][1,2]oxazin-7-one (dPTP) (Zaccolo & Gherardi, 1999). Using plasmid pZE-gfp(ASV) kindly provided by M. Elowitz as template (Elowitz & Leibler, 2000) along with the primers PL_sense_AatII and PL_anti_EcoRI, 10 and 30 amplification cycles with the primers mentioned above were performed. The 151 bp PCR products were purified using the GeneClean Spin Kit (Qbiogene). Following digestion, the product was ligated overnight at 16°C overnight and transformed into library efficiency *E. coli* DH5α (Invitrogen). About 30,000 colonies were screened by eye from minimal media-casamino acid agar plates and 200 colonies, spanning a wide range in fluorescent intensity, were picked from each plate.

10.6.3 Library characterization

10.6.3.1 Initial characterization

About 20 μL of overnight cultures of library clones growing LB broth were used to inoculate 5mL M9G medium supplemented with 0.1% w/v casamino acid (M9G/CAA) and the cultures were grown at 37 °C with orbital shaking. After 14 h, a sample of the culture was centrifuged at 18,000 × *g* for 2 minutes and the cells were resuspended in ice-cold water. Flow cytometry was performed on a Becton-Dickinson FACScan and the

geometric mean of the fluorescence distribution of each clonal population was calculated. In order to ensure that bulk, population-averaged measurements could reflect the underlying single-cell behavior, only clones with clean, monovariate distributions of fluorescence were retained for further analysis. Twenty-seven clones were isolated in this way. Sequencing revealed that these 27 clones represented 22 unique promoter sequences.

10.6.3.2 Promoter strength metric

Shake flasks containing 50 mL of M9G/CAA medium were inoculated with 1% v/v of an overnight LB culture of a library clone. The culture turbidity (A_{600nm}) and fluorescence (Packard Fusion microplate fluorescence reader, Perkin-Elmer, Boston, MA) were monitored as a function of time. Fluorescence readings taken during the exponential growth phase were plotted as a function of turbidity. The best-fit slope to this line represents the exponential-phase steady-state concentration of GFP, f_{SS}. Because f_{SS} is affected by the cell growth rate, oxygen-dependent maturation constant of GFP, and the protease-mediated degradation of GFP as well as the promoter-driven synthesis of new GFP, it is not a suitable metric for promoter strength. Instead, we used a previously published dynamic model (Leveau & Lindow, 2001) that accounts for all of these factors. Estimates of m and D of 1.5 h^{-1} and 0.23 h^{-1}, respectively (Andersen et al., 1998; Cormack, Valdivia, & Falkow, 1996), were obtained from the literature. The parameters f_{SS} and μ were measured separately for each member of the promoter library. P, in relative fluorescence units per absorbance unit per hour, was calculated for each clone. We performed duplicate cultures for each clone.

10.6.3.3 Transcriptional analysis

Cultures inoculated as previously were grown for 3 h and the total RNA was extracted from a 1.5 mL sample with a commercial kit (RNEasy, Qiagen Corp). All samples were diluted to a final concentration of 20 μg/mL and stored at -20 °C. A commercial kit for RT-PCR (iScript One-Step RT-PCR Kit with SYBR Green, Bio-Rad) was used with a CCD-equipped thermal cycler (iCycler, Bio-Rad) for RT-PCR of the *gfp* transcript. Primers were used at a final concentration of 100 nM and 20 ng of RNA was used as template in each 50 μL reaction. We performed duplicate cultures for each clone and duplicate extractions for each culture. The threshold cycles for each sample were calculated from the fluorescence data with proprietary software (Bio-Rad, Inc).

10.6.3.4 Chloramphenicol resistance

pZE-promoter-*cat* plasmids were created by PCR of the CAT gene from pACYC184 using primers CAT_Sense_MluI and CAT_Anti_KpnI and ligated into the proper pZE-promoter construct which was previously digested by KpnI and MluI. Exponential-phase cultures grown in LB supplemented with kanamycin were plated onto LB agar supplemented with kanamycin and various concentrations of chloramphenicol ranging from 0 to 500 μg/ml. After overnight incubation at 37 °C, the lowest concentration of chloramphenicol that inhibited the growth of a clone was recorded.

10.6.4 Promoter delivery

Promoter replacements were conducted using PCR product recombination (Datsenko & Wanner, 2000) using the pKD46 plasmid expressing the lambda red recombination system and pKD13 as the template for PCR. Promoter replacements were

verified through colony PCR using the k1, k2 and kt primers along with the verification primers listed below. To create the cassette for promoter replacement, two fragments were amplified via PCR. Fragment 1 contained the promoter with primer homology to the upstream region of the endogenous promoter. Fragment 2 contained the kanamycin maker from pKD13 and had homology to an area downstream of the endogenous promoter or gene. These two fragments had an internal homology to each other of 25 basepairs to allow for self-annealing and subsequent amplification of a single cassette which was used (~100 ng) for the transformation. For the case of *dxs*, the entire gene was amplified and used as a third fragment which was annealed with the previous two. This provided higher recombination efficiency due to the increased homology region.

10.6.5 List of primers

PL_sense_AatII: TCCGACGTCTAAGAAACCATTATTATC
PL_anti_EcoRI: CCGGAATTCGGTCAGTGCGTCCTGCTGAT

RT-PCR_Sense: ATGGCTAGCAAAGGAGAAGA
RT-PCR_Anti: ATCCATGCCATGTGTAATCC

CAT_Sense_MluI: CGACGCGTATTTCTGCCATTCATCCGCTTATTATCA
CAT_Anti_KpnI: CGGGGTACCTTTCAGGAGCTAAGGAAGCTAAAATGGA

Integration Cassettes

ppc fragment
ppc-pze Sense:
GTTTGATAGCCCTGTATCCTTCACGTCGCATTGGCGCGAATATGCTCGGCATC
TTCCTTTCTCCTCTTTAATGAATTCGG
pze-pkd13 shunt:
GAAGCAGCTCCAGCCTACACTCCGACGTCTAAGAAACCATTATTA
pkd13 sense: GTGTAGGCTGGAGCTGCTTC
pkd13-ppc anti:
CATTTCCATAAGTTACGCTTATTTAAAGCGTCGTGAATTTAATGACGTAATCC
GTCGACCTGCAGTTCGA
verification: CCGATCCCTGGCTATGAATGC

dxs fragment
dxs-pze Sense:
TGGGTGGAGTCGACCAGTGCCAGGGTCGGGTATTTGGCAATATCAAAACTCA
TCACTCCTCTTTAATGAATTCGG
pze-pkd13 shunt:
GAAGCAGCTCCAGCCTACACTCCGACGTCTAAGAAACCATTATTA
pkd13 sense: GTGTAGGCTGGAGCTGCTTC
pkd13-dxs anti:
ACTCGATACCTCGGCACTGGAAGCGCTAGCGGACTACATCATCCAGCGTAAT
AAAATCCGTCGACCTGCAGTTCGA
dxs sense: ATGAGTTTTGATATTGCCAAA

Promoters were sequenced using primers
PL_Left_seq:AGATCCTTGGCGGCAAGAAA and
PL_Right_seq:GCCATGGAACAGGTAGTTTTCCAG

10.7 global Transcription Machinery Engineering

10.7.1 Strains and media

E. coli DH5α (Invitrogen) was used for routine transformations as described in the protocol as well as for all phenotype analysis in this experiment. Strains were grown at 37°C with 225 RPM orbital shaking in either LB-Miller medium or M9-minimal medium containing 5 g/L D-glucose and supplemented with 1mM thiamine (Maniatis, 1982). Media was supplemented with 34 μg/ml of chloramphenicol for low copy plasmid propagation and 68 μg/ml of chloramphenicol, 20 μg/ml kanamycin, and 100 μg/ml ampicillin for higher copy plasmid maintenance as necessary. Cell density was monitored spectrophotometrically at 600 nm. M9 Minimal salts were purchased from US Biological, X-gal was purchased from American Bioanalytical and all remaining chemicals were from Sigma-Aldrich. Primers were purchased from Invitrogen.

10.7.2 Library construction

A low copy host plasmid (pHACM) was constructed using pUC19 (Yanisch-Perron, Vieira, & Messing, 1985) as a host background strain and replacing ampicillin resistance with chloramphenicol using the CAT gene in pACYC184 (Chang & Cohen, 1978) and the pSC101 origin of replication from pSC101 (Bernardi & Bernardi, 1984). The chloramphenicol gene from pACYC184 was amplified with AatII and AhdI restriction site overhangs using primers CM_sense_AhdI: GTTGCCTGACTCCCCGTCGCCAGGCGTTTAAGGGCACCAATAAC and CM_anti_AatII: CAGAAGCCACTGGAGCACCTCAAAACTGCAGT. This fragment was digested along with the pUC19 backbone and ligated together to form pUC19-Cm. The pSC101 fragment from pSC101 was amplified with AflIII and NotI restriction site overhangs using primers pSC_sense_AflIII: CCCACATGTCCTAGACCTAGCTGCAGGTCGAGGA and pSC_anti_NotI: AAGGAAAAAAGCGGCCGCACGGGTAAGCCTGTTGATGATACCGCTGCCTTACT. This fragment was digested along with the pUC19-Cm construct and ligated together to form pHACM.

The *rpoD* gene was amplified from *E. coli* genomic DNA using HindIII and SacI restriction overhangs to target the *lacZ* gene in pHACM to allow for blue/white screening using primers rpoD_sense_SacI: AACCTAGGAGCTCTGATTTAACGGCTTAAGTGCCGAAGAGC and rpoD_anti_HindIII: TGGAAGCTTTAACGCCTGATCCGGCCTACCGATTAAT. Fragment mutagenesis was performed using the GenemorphII Random Mutagenesis kit (Stratagene) using various concentrations of initial template to obtain low, medium, and

high mutation rates as described in the product protocol. Following PCR, these fragments were purified using a Qiagen PCR cleanup kit, digested by HindIII and SacI overnight, ligated overnight into a digested pHACM backbone, and transformed into *E. coli* DH5α competent cells. Cells were plated on LB-agar plates and scraped off to create a liquid library. The total library size of white colonies was approximately 10^6.

10.7.3 Sequence analysis

Sequences of mutant sigma factors were sequenced using the following set of primers:

S1: CCATATGCGGTGTGAAATACCGC, S2: CACAGCTGAAACTTCTTGTCACCC, S3: TTGTTGACCCGAACGCAGAAGA, S4: AGAAACCGGCCTGACCATCG, A1: GCTTCGATCTGACGGATACGTTCG, A2: CAGGTTGCGTAGGTGGAGAACTTG, A3: GTGACTGCGACCTTTCGCTTTG, A4: CATCAGATCATCGGCATCCG, A5: GCTTCGGCAGCATCTTCGT, and A6: CGGAAGCGATCACCTATCTGC. Sequences were aligned and compared using Clustal W version 1.82.

10.7.4 Transcriptional analysis

Ethanol strains were grown to an OD of approximately 0.4 - 0.5 and RNA was extracted using the Qiagen RNeasy Mini Kit. Microarray services were provided by Ambion, Inc. using the Affymetrix E. coli 2.0 arrays. Arrays were run in triplicate with biological replicates to allow for statistical confidence in differential gene expression.

10.7.5 Phenotype selection

Samples from the liquid library were placed into challenging environments to select for surviving mutants. For ethanol tolerance, strains were placed in filtered-LB containing 50 g/L of ethanol. These cultures were performed in 30 x 115 mm closed top centrifuge tubes shaking at 37°C. Strains were plated after 20 hours and selected for individual colony testing. Subsequent round mutants were selected in a similar manner except the selection pressure was increased to 60 or 70 g/L and culture samples were plated after 4 and 8 hours to ensure viability. For lycopene production, strains were selected and cultured as previously reported (Alper et al., 2005b; Alper, Miyaoku, & Stephanopoulos, 2005). SDS and ethanol multiple tolerance mutants were selected in 1% SDS and/or 50 g/L ethanol where appropriate. For acetate tolerance, strains were serial subcultured twice in increasing concentrations of sodium acetate starting at 20 g/L and increasing to 30 g/L in M9 minimal media. Cells were then plated onto LB plates and several colonies were selected for single-colony assays. Routinely, assays for these strains were conducted in M9 minimal medium at 0, 10, 20, and 30 g/L of sodium acetate in addition to 5 g/L glucose with a starting OD600 of 0.04. For pHBA tolerance, strains were cultured in 20 g/L of pHBA in M9 minimal media and plated after 20 hours to select for surviving cells. Routinely, assays for these strains were conducted in M9 minimal medium at 0, 5, 10, and 15 g/L of pHBA with a starting OD600 of 0.02. For n-hexane tolerance, an exponentially phase growing strain of the liquid library was transferred to a 10% hexane/LB medium culture to obtain a hexane saturated culture. These cultivations were performed in 10 ml closed, screw-top glass vials placed into a

shaker at 250 RPM. OD600 was measured after 17 hours by removing a sample from the LB-phase. The plasmids from all strains identified with improved phenotypes were recovered using a Qiagen MiniPrep kit and retransformed into a fresh batch of competent cells.

10.7.6 Yeast examples

S. cerevisiae strain BY4741 (*MAT*a; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*) used in this study was obtained from EUROSCARF, Frankfurt, Germany. It was cultivated in YPD medium (10 g of yeast extract/liter, 20 g of Bacto Peptone/liter and 20 g glucose/liter). For yeast transformation, the Frozen-EZ Yeast Transformation II (ZYMO RESEARCH) was used. To select and grow yeast transformants bearing plasmids with *URA3* as selectable marker, a yeast synthetic complete (YSC) medium was used containing 6.7 g of Yeast Nitrogen Base (Difco)/liter, 20 g glucose/liter and a mixture of appropriate nucleotides and amino acids (CSM-URA, Qbiogene) referred here as to YSC Ura$^-$. Medium was supplemented with 1.5% agar for solid media.

The library was created and cloned behind the TEF-mut2 promoter created previously as part of a yeast promoter library (Alper et al., 2005a). The Taf25 gene was cloned from genomic DNA using the primers TAF25_Sense: TCGAGTGCTAGCAAAATGGATTTTGAGGAAGATTACGAT and TAF25_Anti: CTAGCGGTCGACCTAACGATAAAAGTCTGGGCGACCT. The Spt15 gene was cloned from genomic DNA using the primers SPT15_Sense: TCGAGTGCTAGCAAAATGGCCGATGAGGAACGTTTAAAGG and SPT15_Anti: CTAGCGGTCGACTCACATTTTTCTAAATTCACTTAGCACA. Genes were mutated using the GeneMorph II Mutagenesis Kit and products were digested using NheI and SalI

and ligated to plasmid backbone digested with XbaI and SalI. The plasmids were transformed into *E. coli* DH5α, isolated using a plasmid MiniPrep Spin Kit and transformed into yeast. Plasmids were sequenced using the primers: Seq_Forward: TCACTCAGTAGAACGGGAGC and Seq_Reverse: AATAGGGACCTAGACTTCAG.

Strains were isolated by serial subculturing in 200 to 400 mM LiCl, 200 to 300 g/L of glucose, and 5% Ethanol/100 g/L glucose to 6% Ethanol/120 g/L glucose as appropriate. Cells were isolated by plating onto selective medium plates and assayed for performance. Plasmids were isolated and retransformed to revalidate phenotypes in biological replicates.

Chapter 11

References

Abe, S., et al. (2003). n-Hexane sensitivity of *Escherichia coli* due to low expression of imp/ostA encoding an 87 kDa minor protein associated with the outer membrane. *Microbiology*, 149(Pt 5), 1265-1273.

Adam, P., et al. (2002). Biosynthesis of terpenes: Studies on 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase. *Proc Natl Acad Sci U S A*, 99(19), 12108-12113.

Alper, H., et al. (2005a). Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A*, 102(36), 12678-12683.

Alper, H., et al. (2005b). Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng*, 7(3), 155-164.

Alper, H., Miyaoku, K., & Stephanopoulos, G. (2005). Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol*, 23(5), 612-616.

Andersen, J.B., et al. (1998). New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl Environ Microbiol*, 64(6), 2240-2246.

Aono, R., Negishi, T., & Nakajima, H. (1994). Cloning of organic solvent tolerance gene ostA that determines n-hexane tolerance level in *Escherichia coli. Appl Environ Microbiol*, 60(12), 4624-4626.

Arndt, K.M., et al. (1992). Biochemical and genetic characterization of a yeast TFIID mutant that alters transcription in vivo and DNA binding in vitro. *Mol Cell Biol*, 12(5), 2372-2382.

Askenazi, M., et al. (2003). Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat Biotechnol*, 21(2), 150-156.

Au, D.C., Lorence, R.M., & Gennis, R.B. (1985). Isolation and characterization of an *Escherichia coli* mutant lacking the cytochrome o terminal oxidase. *J Bacteriol.*, 161(1), 123–127.

Badarinarayana, V., et al. (2001). Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotech.*, 19(11), 1060 - 1065.

Bafrncová, P., et al. (1999). Improvement of very high gravity ethanol fermentation by media supplementation using *Saccharomyces cerevisiae. Biotechnology Letters*, 21(4), 337 - 341.

Bai, F.W., et al. (2004). Continuous ethanol production and evaluation of yeast cell lysis and viability loss under very high gravity medium conditions. *J Biotechnol*, 110(3), 287-293.

Bailey, J. (1999). Lessons from Metabolic Engineering for Functional Genomics and Drug Discovery. *Nat Biotechnol*, 17, 616-618.

Bailey, J.E., et al. (2002). Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes. *Biotechnol Bioeng*, 79(5), 568-579.

Barker, J.L. & Frost, J.W. (2001). Microbial synthesis of p-hydroxybenzoic acid from glucose. *Biotechnol Bioeng*, 76(4), 376-390.

Beard, D., Liang, S., & Qian, H. (2002). Energy Balance for analysis of complex metabolic networks. *Biophysics Journal*, 83, 79-86.

Becker, J. & Boles, E. (2003). A Modified *Saccharomyces cerevisiae* Strain That Consumes L-Arabinose and Produces Ethanol. *Appl. Environ. Microbiol.*, 69(7), 4144-4150.

Becker-Hapak, M., et al. (1997). *RpoS* Dependent Overexpression of Carotenoids from *Erwinia herbicola* in OXYR-Deficient *Escherichia coli. Biochemical and Biophysical Research Communications*, 239(1 SU -), 305-309.

Bernardi, A. & Bernardi, F. (1984). Complete sequence of pSC101. *Nucleic Acids Res*, 12(24), 9415-9426.

Berrios-Rivera, S.J., Bennett, G.N., & San, K.-Y. (2002). The Effect of Increasing NADH Availability on the Redistribution of Metabolic Fluxes in *Escherichia coli* Chemostat Cultures. *Metab Eng*, 4(3), 230-237.

Bewley, C.A., Gronenborn, A.M., & Clore, G.M. (1998). Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu Rev Biophys Biomol Struct*, 27, 105-131.

Blake, W.J., et al. (2003). Noise in eukaryotic gene expression. 422(6932), 633-637.

Boder, E.T., Midelfort, K.S., & Wittrup, K.D. (2000). Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci U S A*, 97(20), 10701-10705.

Browning, D.F. & Busby, S.J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1), 57-65.

Burgard, A., Pharkya, P., & Maranas, C.D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotech. Bioeng.*, 84(6), 647-657.

Burgess, R.R. & Anthony, L. (2001). How Sigma Docks to RNA Polymerase and What Sigma Does. *Curr. Opin. Microbiol*, 4, 126-131.

Cameron, D.C., et al. (1998). Metabolic Engineering of Propanediol Pathways. *Biotechnology Progress*, 14(1), 116-125.

Campos, N., et al. (2001). *Escherichia coli* engineered to synthesize isopentenyl diphosphate and dimethylallyl diphosphate from mevalonate: a novel system for the genetic analysis of the 2-C-methyl-d-erythritol 4-phosphate pathway for isoprenoid biosynthesis. *Biochem J*, 353(Pt 1), 59-67.

Chang, A.C. & Cohen, S.N. (1978). Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J Bacteriol*, 134(3), 1141-1156.

Chasman, D.I., et al. (1993). Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proc Natl Acad Sci U S A*, 90(17), 8174-8178.

Chen, R., Hatzimanikatis, V., Yap, W., Postma, P.W. and Bailey, J.E. (1997). Metabolic consequences of Phosphotransferase (PTS) mutation in a phenylalanine-producing recombinant *Escherichia coli. Biotechnology Progress*, 13(6), 768 -775.

Colon, G.E., et al. (1995). Production of isoleucine by overexpression of *ilvA* in a *Corynebacterium lactofermentum* threonine producer. *Appl Microbiol Biotechnol*, 43(3), 482-488.

Cormack, B.P., Valdivia, R.H., & Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 173(1 Spec No), 33-38.

Cunningham, F.X. & Gantt, E. (1998). GENES AND ENZYMES OF CAROTENOID BIOSYNTHESIS IN PLANTS. *Annual Review of Plant Physiology and Plant Molecular Biology*, 49(1), 557-583.

Cunningham FX Jr, S.Z., Chamovitz D, Hirschberg J, Gantt E. (1994). Molecular structure and enzymatic function of lycopene cyclase from the cyanobacterium *Synechococcus sp* strain PCC7942. *Plant Cell*, 6(8), 1107–1121.

Datsenko, K.A. & Wanner, B.L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *PNAS*, 97(12), 6640-6645.

Edwards, J.S. & Palsson, B. (2000). The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA*, 97, 5528-5533.

Edwards, J.S., Ibarra, R.U., & Palsson, B.O. (2001). In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2), 125-130.

Eisen, M., et al. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci USA*, 95, 14863-14868.

Elowitz, M.B. & Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767), 335-338.

Fa, M., et al. (2004). Expanding the substrate repertoire of a DNA polymerase by directed evolution. *J Am Chem Soc*, 126(6), 1748-1754.

Famili, I., et al. (2003). Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A*, 100(23), 13134-13139.

Farmer, W.R. & Liao, J.C. (2000). Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat Biotechnol*, 18(5), 533-537.

Farmer, W.R. & Liao, J.C. (2001). Precursor Balancing for Metabolic Engineering of Lycopene Production in *Escherichia coli*. *Biotechnology Progress*, 17(1), 57-61.

Fell, D.A. (1998). Increasing the flux in metabolic pathways: A metabolic control analysis perspective. *Biotech Bioeng*, 58(2-3), 121–124.

Ferguson, A., et al. (2004). A novel strategy for selection of allosteric ribozymes yields RiboReporter sensors for caffeine and aspartame. *Nucleic Acids Res*, 32(5), 1756-1766.

Forster, J., et al. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res*, 13(2), 244-253.

Fraser, H.I., Kvaratskhelia, M., & White, M.F. (1999). The two analogous phosphoglycerate mutases of *Escherichia coli*. *FEBS Letters*, 455, 344-348.

Gardella, T., Moyle, H., & Susskind, M.M. (1989). A mutant *Escherichia coli* s70 subunit of RNA polymerase with altered promoter specificity. *J. Mol. Biol.*, 206, 579–590.

Gardner, T.S., Cantor, C.R., & Collins, J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli. Nature*, 403(6767), 339-342.

Gerber, H.P., et al. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science*, 263(5148), 808-811.

Gill, R.T., et al. (2002). Genome-wide screening for trait conferring genes using DNA microarrays. *Proc Natl Acad Sci U S A*, 99(10), 7033-7038.

Gill, R.T. (2003). Enabling inverse metabolic engineering through genomics. *Curr Opin Biotechnol*, 14(5), 484-490.

Glieder, A., Farinas, E.T., & Arnold, F.H. (2002). Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nat Biotechnol*, 20(11), 1135-1139.

Gonzalez, R., et al. (2003). Gene array-based identification of changes that contribute to ethanol tolerance in ethanologenic *Escherichia coli*: comparison of KO11 (parent) to LY01 (resistant mutant). *Biotechnol Prog*, 19(2), 612-623.

Gruber, T.M. & Gross, C.A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol*, 57, 441-466.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol*, 11(5), 394-403.

Haro, R., Garciadeblas, B., & Rodriguez-Navarro, A. (1991). A novel P-type ATPase from yeast involved in sodium transport. *FEBS Lett*, 291(2), 189-191.

Hayes, F. (2003). Transposon-Based Strategies for Microbial Functional Genomics and Proteomics. *Annual Review of Genetics*, 37(1), 3-29.

Hecht, S., et al. (2001). Studies on the nonmevalonate pathway to terpenes: The role of the *GcpE* (*IspG*) protein. *Proc Natl Acad Sci U S A*, 98(26), 14837-14842.

Hemmi, H., et al. (1998). Identification of genes affecting lycopene formation in Escherichia coli transformed with carotenoid biosynthetic genes: candidates for early genes in isoprenoid biosynthesis. *J Biochem (Tokyo)*, 123(6), 1088-1096.

Horn, G.T. & Wells, R.D. (1981). The leftward promoter of bacteriophage lambda. Structure, biological activity, and influence by adjacent regions. *J Biol Chem*, 256(4), 2003-2009.

Huang, Q., et al. (2001). Engineering *Escherichia coli* for the Synthesis of Taxadiene, a Key Intermediate in the Biosynthesis of Taxol. *Bioorganic & Medicinal Chemistry*, 9, 2237-2242.

Hutchison, C.A., III, et al. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science*, 286(5447), 2165-2169.

Jana, S. & Deb, J.K. (2005). Strategies for efficient production of heterologous proteins in *Escherichia coli. Appl Microbiol Biotechnol.*

Jensen, P.R. & Hammer, K. (1998). The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Appl Environ Microbiol*, 64(1), 82-87.

Jones, K.L., Kim, S.-W., & Keasling, J.D. (2000). Low-Copy Plasmids can Perform as Well as or Better Than High-Copy Plasmids for Metabolic Engineering of Bacteria. *Metab Eng*, 2(4), 328-338.

Jordan, S.W. & Cronan, J.E., Jr. (2003). The Escherichia coli *lipB* gene encodes lipoyl (octanoyl)-acyl carrier protein:protein transferase. *J Bacteriol*, 185(5), 1582-1589.

Jorgensen, C.M., et al. (2004). Expression of the *pyrG* gene determines the pool sizes of CTP and dCTP in Lactococcus lactis. *Eur J Biochem*, 271(12), 2438-2445.

Kacser, H. & Acerenza, L. (1993). A universal method for achieving increases in metabolite production. *Eur J Biochem*, 216(2), 361-367.

Kajiwara, S., et al. (1997). Expression of an exogenous isopentyl diphosphate isomerase gene enhances isoprenoid biosynthesis in *Escherichia coli. Biochem. J.*, 324(2), 421-426.

Kang, M.J., et al. (2005). Identification of genes affecting lycopene accumulation in Escherichia coli using a shot-gun method. *Biotechnol Bioeng*, 91(5), 636-642.

Kauffman, K.J., Prakash, P., & Edwards, J. (2003). Advances in Flux Balance Analysis. *Current Opinion in Biotechnology*, 14(5), 491-496.

Khlebnikov, A., et al. (2000). Regulatable arabinose-inducible gene expression system with consistent control in all cells of a culture. *J Bacteriol*, 182(24), 7029-7034.

Khosla, C. & Bailey, J.E. (1988). Heterologous expression of a bacterial haemoglobin improves the growth properties of recombinant *Escherichia coli. Nature*, 331, 633-635.

Kim, J. & Iyer, V.R. (2004). Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol Cell Biol*, 24(18), 8104-8112.

Kim, J.L., Nikolov, D.B., & Burley, S.K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365(6446), 520-527.

Kim, J.S., et al. (1997). Design of TATA box-binding protein/zinc finger fusions for targeted regulation of gene expression. *Proc Natl Acad Sci U S A*, 94(8), 3616-3620.

Kim, S.-W. & Keasling, J.D. (2001). Metabolic engineering of the nonmevalonate isopentenyl diphosphate synthesis pathway in *Escherichia coli* enhances lycopene production. *Biotechnology and Bioengineering*, 72(4), 408-415.

Kirchner, J., et al. (2001). Molecular genetic dissection of TAF25, an essential yeast gene encoding a subunit shared by TFIID and SAGA multiprotein transcription factors. *Mol Cell Biol*, 21(19), 6668-6680.

Koffas, M.A.G., Jung, G.Y., & Stephanopoulos, G. (2003). Engineering metabolism and product formation in *Corynebacterium glutamicum* by coordinated gene overexpression. *Metabolic Engineering*, 5(1), 32-41.

Kou, H., et al. (2003). Structural and functional analysis of mutations along the crystallographic dimer interface of the yeast TATA binding protein. *Mol Cell Biol*, 23(9), 3186-3201.

Lasko, D.R., et al. (1997). Acetate-specific stress response in acetate-resistant bacteria: an analysis of protein patterns. *Biotechnol Prog*, 13(5), 519-523.

Lasko, D.R., Zamboni, N., & Sauer, U. (2000). Bacterial response to acetate challenge: a comparison of tolerance among species. *Appl Microbiol Biotechnol*, 54(2), 243-247.

Lee, J.H., Van Montagu, M., & Verbruggen, N. (1999). A highly conserved kinase is an essential component for stress tolerance in yeast and plant cells. *Proc Natl Acad Sci U S A*, 96(10), 5873-5877.

Lee, J.-Y., Roh, J.-R., & Kim, H.-S. (1994). Metabolic Engineering of *Pseudomonas putida* for the simultaneous biodegradation of benzene, toluene, and p-xylene mixture. *Biotech Bioeng*, 43, 1146-1152.

Lee, K., et al. (2003). Profiling of dynamic changes in hypermetabolic livers. *Biotechnology and Bioengineering*, 83(4), 400-415.

Lee, P.C. & Schmidt-Dannert, C. (2002). Metabolic engineering towards biotechnological production of carotenoids in microorganisms. *Applied Microbiology and Biotechnology*, 60(1), 1-11.

Lee, T.I. & Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34, 77-137.

Leinfelder, W., et al. (1988). *Escherichia coli* genes whose products are involved in selenium metabolism. *J Bacteriol*, 170(2), 540-546.

Leveau, J.H. & Lindow, S.E. (2001). Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J Bacteriol*, 183(23), 6752-6762.

Lewis, B., et al. (2003). Prediction of mammalian microRNA targets. *Cell*, 115(7), 787-798.

Liao, J.C., Chao, Y.P., & Patnaik, R. (1994). Alteration of the biochemical valves in the central metabolism of *Escherichia coli*. *Ann N Y Acad Sci*, 745, 21-34.

Liu, Y.-G. & Whittier, R.F. (1995). Thermal Asymmetric Interlaced PCR: Automatable Amplification and Sequencing of Insert End Fragments from PI and YAC Clones for Chromosome Walking. *Genomics*, 25(3), 674-681.

Lonetto, M.A., et al. (1998). Identification of a contact site for different transcription activators in region 4 of the *Escherichia coli* RNA polymerase sigma70 subunit. *J Mol Biol*, 284(5), 1353-1365.

Lutz, R. & Bujard, H. (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res*, 25(6), 1203-1210.

Malhotra, A., Severinova, E., & Darst, S.A. (1996). Crystal structure of a sigma 70 subunit fragment from *E. coli* RNA polymerase. *Cell*, 87(1), 127-136.

Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Martin, V.J.J., et al. (2003). Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature Biotechnology*, 21(7), 796 - 802.

Masse, E. & Gottesman, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 99(7), 4620-4625.

Mathews, P.D. & Wurtzel, E.T. (2000). Metabolic engineering of carotenoid accumulation in *Escherichia coli* by modulation of the isoprenoid precursor pool with expression of deoxyxylulose phosphate synthase. *Applied Micro Biotech*, 53(4), 396-400.

McAlister, L.E., Evans, E.L., & Smith, T.E. (1981). Properties of a mutant *Escherichia coli* phosphoenolpyruvate carboxylase deficient in coregulation by intermediary metabolites. *J Bacteriol*, 146(1), 200-208.

Miller, J.H. (1992). *A Short Course in Bacterial Genetics*. Cold Springs Harbor, NY: Cold Springs Harbor Laboratory Press.

Mnaimneh, S., et al. (2004). Exploration of essential gene functions via titratable promoter alleles. *Cell*, 118(1), 31-44.

Muffler, A., et al. (1996). The response regulator *RssB* controls stability of the sigma(S) subunit of RNA polymerase in *Escherichia coli*. *EMBO J.*, 15(6), 1333-1339.

Nelms, J., et al. (1992). Novel mutations in the pheA gene of *Escherichia coli* K-12 which result in highly feedback inhibition-resistant variants of chorismate mutase/prephenate dehydratase. *Appl Environ Microbiol*, 58(8), 2592-2598.

Niederberger, P., et al. (1992). A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. *Biochem J*, 287((Pt 2)), 473-479.

Nishino, K., Inazumi, Y., & Yamaguchi, A. (2003). Global analysis of genes regulated by *EvgA* of the two-component regulatory system in *Escherichia coli*. *J Bacteriol*, 185(8), 2667-2672.

Ostergaard, S., et al. (2000). Increasing galactose consumption by *Saccharomyces cerevisiae* through metabolic engineering of the GAL gene regulatory network. *Nat Biotech*, 18(12), 1283 - 1286.

Owens, J.T., et al. (1998). Mapping the sigma 70 subunit contact sites on *Escherichia coli* RNA polymerase with a sigma 70-conjugated chemical protease. *PNAS*, 95(11), 6021-6026.

Park, K.S., et al. (2003). Phenotypic alteration of eukaryotic cells using randomized libraries of artificial transcription factors. *Nat Biotechnol*, 21(10), 1208-1214.

Prieto, M., Diaz, E., & Garcia, J. (1996). Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of *Escherichia coli* W: engineering a mobile aromatic degradative cluster. *J. Bacteriol.*, 178(1), 111-120.

Rawlings, N., Tolle, D., & Barrett, A. (2004). MEROPS: the peptidase database. *Nucleic Acids Research*, 32, D160-D164.

Reed, J., et al. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology*, 4(9), R54.

Riesenberg, D., et al. (1991). High cell density cultivation of *Escherichia coli* at controlled specific growth rate. *J Biotechnol*, 20(1), 17-27.

San, K.Y., et al. (1994). An optimization study of a pH-inducible promoter system for high-level recombinant protein production in *Escherichia coli*. *Ann NY Acad Sci*, 721(1), 268-276.

Sandmann, G., Woods, W., & Tuveson, R.W. (1990). Identification of carotenoids in *Erwinia herbicola* and in a transformed *Escherichia coli* strain. *FEMS Microbiol Lett*, 59(1-2), 77-82.

Sandmann, G., et al. (1999). The biotechnological potential and design of novel carotenoids by gene combination in *Escherichia coli*. *Trends in Biotechnology*, 17(6), 233-237.

Sandmann, G. (2002). Combinatorial Biosynthesis of Carotenoids in a Heterologous Host: A Powerful Approach for the Biosynthesis of Novel Structures. *ChemBioChem*, 3(7), 629-635.

Schmidt-Dannert, C., Umeno, D., & Arnold, F.H. (2000). Molecular breeding of carotenoid biosynthetic pathways. *Nat Biotechnol*, 18(7), 750-753.

Schultz, M.C., Reeder, R.H., & Hahn, S. (1992). Variants of the TATA-binding protein can distinguish subsets of RNA polymerase I, II, and III promoters. *Cell*, 69(4), 697-702.

Segre, D., Vitkup, D., & Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23), 15112-15117.

Serres, M.H., et al. (2001). A functional update of the *Escherichia coli* K-12 genome. *Genome Biology*, 2(9), research0035.0031-0035.0037.

Sharma, U.K., et al. (1999). Study of the Interaction between Bacteriophage T4 asiA and *Escherichia coli* sigma 70, Using the Yeast Two-Hybrid System: Neutralization of asiA Toxicity to *E. coli* Cells by Coexpression of a Truncated sigma 70 Fragment. *J. Bacteriol.*, 181(18), 5855-5859.

Shastri, A.A. & Morgan, J.A. (2005). Flux Balance Analysis of Photoautotrophic Metabolism. *Biotechnol. Prog.*

Shimizu, K., et al. (2005). Discovery of glpC, an organic solvent tolerance-related gene in *Escherichia coli*, using gene expression profiles from DNA microarrays. *Appl Environ Microbiol*, 71(2), 1093-1096.

Shlomi, T., Berkman, O., & Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21), 7695-7700.

Siegele, D.A., et al. (1989). Altered promoter recognition by mutant forms of the sigma 70 subunit of *Escherichia coli* RNA polymerase. *J Mol Biol*, 206(4), 591-603.

Siegele, D.A. & Hu, J.C. (1997). Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc Natl Acad Sci U S A*, 94(15), 8168-8172.

Soga, T., et al. (2003). Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry. *Journal of Proteome Research*, 2, 488- 494.

Spencer, J.V. & Arndt, K.M. (2002). A TATA binding protein mutant with increased affinity for DNA directs transcription from a reversed TATA sequence in vivo. *Mol Cell Biol*, 22(24), 8744-8755.

Stafford, D.E., et al. (2002a). Optimizing bioconversion pathways through systems analysis and metabolic engineering. *Proc Natl Acad Sci U S A*, 99(4), 1801-1806.

Stafford, D.E., et al. (2002b). Optimizing bioconversion pathways through systems analysis and metabolic engineering. *PNAS*, 99(4), 1801-1806.

Stemmer, W.P. (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, 370(6488), 389-391.

Stephanopoulos, G. & Vallino, J.J. (1991). Network rigidity and metabolic engineering in metabolite overproduction. *Science*, 252(5013), 1675-1681.

Stephanopoulos, G., Aristidou, A., & Nielsen, J. (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego: Academic Press.

Stephanopoulos, G. (1999). Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering*, 1(1), 1-11.

Stephanopoulos, G. (2002). Metabolic Engineering: Perspective of a Chemical Engineer. *AIChE Journal*, 48(5), 920-926.

Stephanopoulos, G., Alper, H., & Moxley, J. (2004). Exploiting biological complexity for strain improvement through systems biology. *Nat Biotechnol*, 22(10), 1261-1267.

Tao, L., Jackson, R.E., & Cheng, Q. (2005). Directed evolution of copy number of a broad host range plasmid for metabolic engineering. *Metab Eng*, 7(1), 10-17.

Thatipamala, R., Rohani, S., & Hill, G. (1992). Effects of high product and substrate inhibitions on the kinetics and biomass and product yields during ethanol batch fermentation. *Biotechnology and Bioengineering*, 40(2), 289-297.

Umeno, D., Tobias, A.V., & Arnold, F.H. (2002). Evolution of the C30 carotenoid synthase CrtM for function in a C40 pathway. *J Bacteriol*, 184(23), 6690-6699.

Umeno, D., Tobias, A.V., & Arnold, F.H. (2005). Diversifying carotenoid biosynthetic pathways by directed evolution. *Microbiol Mol Biol Rev*, 69(1), 51-78.

Van Dyk, T.K., et al. (2004). Characterization of the *Escherichia coli* AaeAB efflux pump: a metabolic relief valve? *J Bacteriol*, 186(21), 7196-7204.

Wang, C.-W., Oh, M.-K., & Liao, J.C. (1999). Engineered isoprenoid pathway enhances astaxanthin production in *Escherichia coli. Biotechnology and Bioengineering*, 62(2), 235-241.

Watanabe, K., et al. (2003). Engineered biosynthesis of an ansamycin polyketide precursor in *Escherichia coli. Proc Natl Acad Sci U S A*, 100(17), 9774-9778.

Wells, J.A. (1990). Additivity of mutational effects in proteins. *Biochemistry*, 29, 8509–8517.

Wiechert, W. (2002). Modeling and simulation: tools for metabolic engineering. *Journal of Biotechnology*, 94(1), 37-63.

Yanisch-Perron, C., Vieira, J., & Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, 33(1), 103-119.

Yarmush, M.L. & Banta, S. (2003). METABOLIC ENGINEERING: Advances in Modeling and Intervention in Health and Disease. *Annual Review of Biomedical Engineering*, 5(1), 349-381.

Yomano, L.P., York, S.W., & Ingram, L.O. (1998). Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. *J Ind Microbiol Biotechnol*, 20(2), 132-138.

Zaccolo, M., et al. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J Mol Biol*, 255(4), 589-603.

Zaccolo, M. & Gherardi, E. (1999). The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 beta-lactamase. *J Mol Biol*, 285(2), 775-783.

Zaldivar, J., Nielsen, J., & Olsson, L. (2001). Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. *Applied Micro Biotech*, 56(1 - 2), 17-34.

Zaslaver, A., et al. (2004). Just-in-time transcription program in metabolic pathways. *Nat Genet*, 36(5), 486-491.

Zhang, X., et al. (1995). Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng*, 8, 1017–1022.

Zhou, L., et al. (2003). Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems. *J Bacteriol*, 185(16), 4956-4972.

Zinoni, F., et al. (1986). Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. *Proc Natl Acad Sci U S A*, 83(13), 4650-4654.