Genomic engineering of Escherichia coli for improved ethanol tolerance and ethanol production

Lauren B. A. Woodruff
University of Colorado at Boulder, lauren.b.woodruff@gmail.edu

Follow this and additional works at: https://scholar.colorado.edu/chbe_gradetds

Recommended Citation
https://scholar.colorado.edu/chbe_gradetds/40
GENOMIC ENGINEERING OF *ESCHERICHIA COLI* FOR
IMPROVED ETHANOL TOLERANCE AND ETHANOL PRODUCTION

by

LAUREN B. A. WOODRUFF

B.S., Cornell University, 2006
M.S., University of Colorado, 2009

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Chemical and Biological Engineering
2012
This thesis entitled:

Genomic engineering of *Escherichia coli* for improved ethanol tolerance and ethanol production

written by Lauren B. A. Woodruff

has been approved for the Department of Chemical and Biological Engineering

____________________________

Ryan T. Gill

____________________________

J. Will Medlin

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Woodruff, Lauren B. A. (Ph.D., Chemical and Biological Engineering)

Genomic engineering of *Escherichia coli* for improved ethanol tolerance and ethanol production

Thesis directed by Associate Professor Ryan T. Gill

With tremendous advancements in the genomics era, engineered microbes are being used as biological factories to produce an array of products, including pharmaceuticals, bioplastics, commodity chemicals, and biofuels. The microbial machinery can be tailored through genetic modifications that result in traits which enable these industrial processes, such as new products, nonnative feedstocks, more efficient production, or more robust cellular biocatalysts. Nevertheless, identifying these trait-conferring genetic modifications is often laborious, and the toxicity of the product is a key challenge for commercial production of fuels and chemicals. In this thesis research, we have used the *E. coli* model system to study genome-wide approaches to identify trait-conferring genetic modifications for improved tolerance to and production of ethanol, an industrially-relevant biofuel. We mapped the genome for ethanol tolerance in a wild-type strain using the high resolution multiscalar analysis of library enrichments (SCALEs) approach. By testing the highest fitness genes identified, we confirmed nine novel genetic targets that confer improved ethanol tolerance. Transcriptomic and proteomic analysis of the ethanol stress response in these engineered ethanol-tolerant clones identified expression changes relating to a subset of biological processes shared among these clones. We constructed an *E. coli* ethanol production platform strain containing fully characterized genetic modifications that produces ethanol comparable to the best previously reported, and this strain is openly available from the ATCC. Using the LW06 production platform, we tested ethanol tolerance-conferring
genes we previously identified, and the tolerance-conferring quality of these genes was largely dependent on the system, such as the host genotype, metabolism, and media. Based on this finding, we designed a genome-wide selection for ethanol production using SCALEs and mapped the fitness of the entire genome under ethanol production conditions. We selected for clones with both improved ethanol tolerance and production and through combinatorial testing, identified genes which confer improved ethanol tolerance and increased rates of ethanol production. From the quantitative genome-wide mapping, we propose future work to further engineer complex, multigenic traits for improved *E. coli* ethanol production. Collectively, the work presented here has identified strategies applicable to engineering improved production of inhibitory products.
In loving memory of

Max and Ida Feryszka
Acknowledgements

First off, I would like to give Ryan my utmost gratitude for mentoring me throughout this process, guiding the path that this project taken, consistently challenging me to aim higher and higher, and trusting me with the freedom to own my project, which taught me how to overcome research obstacles and ultimately become a better scientist. And thanks for standing by my work and abilities, offering pep talks when I doubted myself, like a coach on the sidelines.

I thank the members of my committee, Arthi Jayaraman, Barbara Demmig-Adams, Joel Kaar, and Will Medlin, for taking the time to participate in my project and offering helpful advice over the past few years that has helped shape this project. I’d like to especially thank Will for his help and patient guidance in all of the general grad school-related issues that brought me to his door, dating all the way back to my initial recruiting visit. Barbara, I sincerely thank you for urging me to reconsider academia.

To the Gill group past and present, you all have played a huge role in this work, and I thank you guys for all of the help and camaraderie. Joe Warner, I learned so much from you. Thanks for the brainstorming sessions and invaluable lessons (e.g., rigorous primer design, always include controls) that tremendously improved my research. I give a big thanks to Nich Sandoval, Eileen Spindler, Paul Handke, Pip Reeder, Sean Lynch, and Jamie Prior for putting up with all of my endless questioning that helped me think through and figure out various aspects of this work.
I would like to thank my collaborators and teammates that contributed to the publications resulting from this project. Jags Pandhal, Saw Yen Ow, and Phil Wright at the University of Sheffield and also Anis Karimpour-Fard and Sophie Weiss contributed to Chapter 2. Brian May and Joe Warner contributed to Chapter 3. Nanette Boyle contributed to Chapter 4. Brian May, Joe Warner, and Sean Lynch contributed to Chapter 5. I greatly appreciate your help. Funding that supported this research was graciously provided through a Graduate Research Supplement from the U.S. National Science Foundation (CBET0449183) and fellowship assistance from the U.S. Department of Education’s Graduate Assistantships in Areas of National Need program.

To all of my amazing family, friends, and mentors through the years, I really would not be here without you, and I cannot possibly thank you enough for your teaching, encouragement, and assistance along this journey. I have been incredibly lucky to have such great friends to turn to during this time in grad school and throughout my whole life. I’ve learned so much from each and every one of you. To my brother, James, who I’ve always looked up to, thanks for taking me under your wing and never hesitating to lend a hand or an ear. To my parents, thank you for instilling persistence, curiosity, and a love of learning in me from the very beginning. A very heartfelt thanks to my Mom for being my biggest champion in all of my endeavors. To my husband, Woody, thank you for always being there to lean on, truly believing I can accomplish anything, being genuinely interested in and proud of my work (even when it required crazy schedules to complete), pushing me to think outside of the box on anything and everything, and solving my many equipment and technical issues with your mad skills. I can’t imagine a more enjoyable, supportive, and loving companion. To many more adventures and accomplishments. The sky’s the limit.
# Table of Contents

## CHAPTER 1  ENGINEERING GENOMES IN MULTIPLEX

1.1 Abstract .............................................................................................................. 1
1.2 Introduction ......................................................................................................... 1
1.3 A genome engineering algorithm ..................................................................... 2
1.4 Genome search technologies .......................................................................... 6
1.5 Assigning relevance ......................................................................................... 10
1.6 Combinatorial optimization ........................................................................... 11
1.7 Conclusions ....................................................................................................... 14
1.8 References ......................................................................................................... 15

## CHAPTER 2  GENOME-SCALE IDENTIFICATION AND CHARACTERIZATION OF ETHANOL TOLERANCE GENES IN *ESCHERICHIA COLI*

2.1 Abstract .............................................................................................................. 21
2.2 Introduction ......................................................................................................... 22
2.3 Materials and methods ..................................................................................... 24
  2.3.1 Strains, media, and reagents ................................................................. 24
  2.3.2 Genomic library preparation and selections ......................................... 25
  2.3.3 Quantification of selection enrichment by microarray analysis .......... 26
  2.3.4 Gene fitness determination ................................................................. 26
  2.3.5 Clone construction ............................................................................... 27
  2.3.6 Growth testing and ANOVA ............................................................... 28
  2.3.7 Transcriptional profiling ................................................................... 28
  2.3.8 iTRAQ proteomics .......................................................................... 29
  2.3.9 Clustering and gene ontology enrichment analysis ........................ 31
2.4 Results and discussion ...................................................................................... 31
  2.4.1 Genome-wide selections for ethanol tolerance ..................................... 31
  2.4.2 Identification and confirmation of ethanol tolerance genes ................. 35
  2.4.3 Transcriptional and Proteomic analysis of ethanol-tolerant clones ...... 41
2.5 Conclusions ....................................................................................................... 53
2.6 References ........................................................................................................54

CHAPTER 3  TOWARDS A METABOLIC ENGINEERING STRAIN “COMMONS”: AN ESCHERICHIA COLI PLATFORM STRAIN FOR ETHANOL PRODUCTION .......................................................... 59

3.1 Abstract .............................................................................................................. 59
3.2 Introduction ........................................................................................................ 60
3.3 Results and Discussion .................................................................................... 64
3.4 Conclusions ........................................................................................................ 71
3.5 Materials and Methods .................................................................................... 72
  3.5.1 Bacterial strains, plasmids, and media ......................................................... 72
  3.5.2 Genetic methods .......................................................................................... 74
  3.5.3 Fermentations and tolerance testing ........................................................... 77
  3.5.4 Analytical techniques ................................................................................ 78
3.6 References ......................................................................................................... 79

CHAPTER 4  ENGINEERING IMPROVED ETHANOL PRODUCTION IN ESCHERICHIA COLI WITH A GENOME-WIDE APPROACH ........................................................................... 82

4.1 Abstract .............................................................................................................. 82
4.2 Introduction ........................................................................................................ 83
4.3 Materials and methods ..................................................................................... 85
  4.3.1 Strains, media and reagents ....................................................................... 85
  4.3.2 Library construction .................................................................................... 86
  4.3.3 Selections and microarray quantification .................................................... 87
  4.3.4 Ethanol tolerance testing ............................................................................ 89
  4.3.5 Batch fermentations ................................................................................... 90
  4.3.6 Clone construction ..................................................................................... 90
4.4 Results and Discussion .................................................................................... 91
  4.4.1 Designing a directed, genome-wide selection for ethanol production ........ 91
  4.4.2 E. coli genome mapping for ethanol production trait-conferring genotypes .. 94
  4.4.3 Testing ethanol tolerance and ethanol production of selected clones .......... 98
  4.4.4 Combinatorial analysis of ethanol tolerance and ethanol production genes ... 104
<table>
<thead>
<tr>
<th>Chapter Title</th>
<th>Section Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions</td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>108</td>
</tr>
<tr>
<td>Chapter 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Future Work:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Engineering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trait-Specific</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regulators</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Libraries</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of Parts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>to Improve</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E. Coli</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethanol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Production</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>124</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Materials</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and Methods</td>
<td></td>
</tr>
<tr>
<td>5.2.1</td>
<td>Strains</td>
<td></td>
</tr>
<tr>
<td></td>
<td>, media,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and reagents</td>
<td></td>
</tr>
<tr>
<td>5.2.2</td>
<td>Clone</td>
<td></td>
</tr>
<tr>
<td></td>
<td>construction</td>
<td></td>
</tr>
<tr>
<td>5.2.3</td>
<td>Library</td>
<td></td>
</tr>
<tr>
<td></td>
<td>construction</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and Discussion</td>
<td>129</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Identification of trait-specific regulators</td>
<td>129</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Regulon library design and construction</td>
<td>133</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Future work</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>References</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Chapter 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and Future</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Directions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>141</td>
</tr>
<tr>
<td>6.1</td>
<td>References</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>145</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>147</td>
</tr>
</tbody>
</table>
List of Tables

Table 2-1 Primers used to amplify genomic regions containing genes of interest. ...................... 27
Table 2-2 Constructed clones containing highest-fitness genes identified from ethanol tolerance selections ........................................................................................................................................ 38
Table 2-3 Enriched Up GO terms for all clones determined from transcriptional profiling data. List of all GO terms enriched (p< 0.05) for the 1.5-fold overexpressed genes in all clones. 46
Table 2-4 Significantly up- or down-regulated proteins in lpcA clone (p< 0.05) determined by iTRAQ. .................................................................................................................................................. 48
Table 2-5 Significantly up- or down-regulated proteins in tilS clone (p< 0.05) determined by iTRAQ. .................................................................................................................................................. 49
Table 2-6 Upregulated GO terms for lpcA and tilS under ethanol stress from iTRAQ proteomic analysis. List of all GO terms enriched (p< 0.05) for the significantly upregulated proteins. ............................................................................................................................................................................ 52
Table 3-1 Strains and plasmids used in this study. ........................................................................ 73
Table 3-2 PCR reactions and primers used for strain construction ............................................. 76
Table 3-3 Primers used for strain verification .................................................................................. 77
List of Figures

Figure 1-1 A genome engineering algorithm comprised of these three steps allows for an efficient search of the relevant combinatorial genotypic space for a trait of interest. (A) Genetic diversity on the genome-scale is constructed by targeting functional modifications to loci throughout the genome. A screen or selection for the desired trait is used to narrow the genetic diversity and enrich for individual genetic changes. (B) Relevance of the genetic changes is ranked based on results of the screen/selection and additional relevance metrics. Ranking and further narrowing the number of genes targeted directs the combinatorial search to comprehensively search the relevant genotypic space. (C) Genetic diversity targeted to fine-tune the most relevant genetic changes in parallel is introduced. A screen or selection for the desired trait is used to narrow the genetic diversity by enriching for the combinatorially-optimized, complex (multigenic) trait of interest........................................ 5

Figure 1-2 Combinatorial search space. The combinatorial genotypic space (number of possible genotypes comprised of combinations of genetic targets) grows rapidly when additional loci are targeted as depicted by the combinatorial space for (A) 2 loci (B) 4 loci and (C) 6 loci, where each letter represents an individual locus. Figure is adapted from Wright’s original adaptive landscapes (Wright 1932). ........................................................................................................ 12

Figure 2-1 Design of ethanol tolerance selection strategy utilizing the SCALEs approach. Four plasmid libraries containing either 1, 2, 4, or 8 kb fragments of E. coli genomic DNA were constructed and in sufficient numbers to ensure 99.9% probability of entire genome representation in each library. Serial batch selections were carried out with 15 g/L or 30 g/L ethanol addition. Microarrays were used to quantify the population clones at the end of each selection as well as the starting library population. ................................................................. 32

Figure 2-2 Enrichment of genomic library clones during ethanol selections. (a) Chromosomal enrichment and fitness of clones enriched during 15 g/L ethanol selection and 30 g/L ethanol selections. Each peak represents the natural log fitness, ln(W), of a clone containing a genomic fragment corresponding to its location on the chromosomal map (genomic start position at twelve o’clock). Peak color denotes the size of the clone’s genomic fragment. (b) Histogram of the clonal fitness for all unique clones quantified in each selection. ......................... 34

Figure 2-3 Gene fitness and ethanol tolerance of top fitness genes. (a) Fitness of each gene (calculated as the sum of clone fitness values for all clones containing at least 75% of the gene) in each selection. Highest fitness genes from selections are colored and labeled, and clones containing these genes with their native promoters were constructed. (b) Ethanol tolerance of constructed clones containing top genes was measured in a microtiter plate at 16 hours of growth. Error bars for each clone indicate standard deviation for 8 replicate wells on 3 separate plates (n=24). Empty vector control repeated on each plate (n=48). P-values were calculated by a 2-way ANOVA with factors accounting for clone and plate.......... 36
Figure 2-4 Comparison of gene fitness values reported here and previously by Goodarzi, et al (2010). Our MOPS minimal media selections were carried out at ethanol concentrations of 15 g/L (moderate) and 30 g/L (high), whereas Goodarzi et al. (2010) performed selections in LB rich media at 32 g/L (moderate) and 40 g/L ethanol (high).

Figure 2-5 Ethanol tolerance transcriptomes. (a) Transcriptomes of ethanol-tolerant clones. Hierarchical clustering of the ethanol-stressed transcriptomes based on the concentration of each gene’s mRNA concentration (Pearson’s correlation rows and columns and average linkage) shows strong similarity between biological replicates and two main clusters for the clones’ transcriptomes. Figure generated using GenePattern. (b-e) Venn diagrams generated using VennMaster for (b) overexpressed and (c) underexpressed genes (1.5-fold average) for each ethanol-tolerant clone compared to the parent strain. Gene expression changes were mapped to GO terms to evaluate significant (d) enrichment and (e) depletion of GO terms for each ethanol-tolerant clone.

Figure 2-6 Clustering of expression microarrays in ethanol tolerance engineering studies. Hierarchical clustering was performed to compare the altered transcriptomes of the ethanol-tolerant clones engineered in this study to those engineered in previous studies, specifically ethanol-tolerant clones containing mutant rpoD sigma factors engineered by 3 rounds of gTME (Alper et al., 2007 under GEO accession number GSE3665) and ethanol-tolerant clones selected by parallel adaptation to 5% (v/v) (~40 g/L) ethanol for 2,500 hours (Horinouchi et al., 2010 in supplementary table 1). Absent or zero-containing rows were eliminated for this analysis. Clustering was performed over a total of 2,145 genes.

Figure 2-7 Analysis of ethanol-tolerant proteomes. (a) Comparison of genes upregulated and downregulated in the ethanol-stressed transcriptomes and proteomes for two most ethanol-tolerant clones identified (lpcA and tilS) as compared to the parent strain. (b) GO enrichment overlap for GO terms with significant overrepresentation in for the upregulated and downregulated proteins in lpcA and tilS clones.

Figure 2-8 Comparison of expression changes at RNA and protein level for lpcA and tilS clones. The average fold change from transcriptional profiling and proteomics for upregulated protein lists of lpcA clone (blue diamonds) and tilS clone (red squares).

Figure 3-1 Ethanol production chromosomal integration and activity. The ethanol production cassette comprised of pdc and adhB from Z. mobilis under the control the P_{lac} promoter and an ampicillin-resistance marker were integrated into the chromosome at the attTn7, replacing base-pair 3,909,796, by λ Red-mediated homologous recombination using pSIM5 and resulting in strain LW01. The P_{lac} promoter was swapped for the P_{LlacO-1} to achieve higher activity of the ethanol production pathway. Activity of the ethanol production pathway was tested using aldehyde indicator plates (Conway et al., 1987) with 1 mM IPTG induction.

Figure 3-2 Ethanol production of LW02 and further engineered LW06. (A) Batch fermentation of LW02 and (B) fermentative products produced at 24 hr and 96 hr during the fermentation. (C) Batch fermentation of LW06 (LW02 ΔadhE ΔackA ΔfrdABCD ΔldhA) and (D) fermentative products produced at 24 hr and 96 hr during the fermentation. Ethanol (circles) and OD_{600} (triangles) are represented in (A) and (C).
Figure 3-3 Testing previously identified ethanol tolerance genes in an ethanol production system.
(A) Batch fermentations of the LW05 production strain containing the empty vector control (blue, left), the \textit{tilS}-pSM plasmid (yellow, middle), \textit{lpcA}-pSM (red, right) with glucose (squares), ethanol (circles), and \textit{OD}_{600} (triangles) shown for each. (B) Correlation between ethanol produced per unit cell mass (\textit{OD}_{600}) during the batch fermentation time course for all three clones. (C) Tolerance testing in 25 g/L ethanol for the indicated plasmids under varying growth conditions with the host strain and media denoted for each condition. The \textit{OD}_{600} after 24 hr of growth in the denoted condition was measured for triplicate biological replicates.

Figure 4-1 Design of selections for ethanol tolerance and production using SCALEs libraries. (A) The previously constructed LW06 E.coli platform for ethanol production was used as the genomic DNA for the genomic libraries preparation and also as the host strain for the plasmid libraries. (B) SCALEs genomic libraries were constructed by fragmenting purified LW06 genomic DNA by restriction digest (2 four-cutters). The fragmented genomic DNA was size separated and purified bands corresponding to approximately 1, 2, and 4 kb were ligated to the pSMART-LCKan vector backbone to create the plasmid-based genomic libraries. Each ligation reaction was transformed into the LW06 ethanol production host strain, and greater than 106 transformed cells were generated for each library, which is >99.9 % probability of the entire genome being represented in each library. (C) The libraries were mixed together in equal proportions and grown under batch ethanol production conditions in AMX ethanol production media (100 g/L glucose, 1 mM IPTG induction of ethanol production) dosed with 15 g/L ethanol. After each 24 hr of batch growth, the selection culture was serially transferred for a total of 8 batches with identical conditions. Samples were saved from the end of the fourth and eighth batch and the initial mixed library for microarray analysis. Plasmid DNA was extracted from each sample and quantified by microarray analysis. (D) A control selection to test the potential for improvement by natural adaptation under these conditions was performed in parallel under identical conditions, except using the LW06 strain harboring the empty pSMART-LCKan vector (parent) in place of the libraries, but no improvement in growth was observed for adaptation during this selection, whereas the genomic libraries shows improved batch growth at the end of selection.

Figure 4-2 Genome enrichment profiles during ethanol selection. Chromosomal maps showing the clonal fitness (final concentration over the initial concentration for clone i) for all quantified clones at the fourth batch (A) and eighth batch (B) of selection. Each peak on the chromosomal represents a clone; the height of the peak corresponds to the clonal fitness; the color (red=1 kb, yellow=2 kb, green=4 kb) denotes the size the clone’s genomic fragment; and the location around the chromosome (genome nt 1 at 12 o’clock) corresponds to the locus of the genomic fragment. (C) The fitness for each gene at each quantified selection time point was quantified by summing the clonal fitness values for all clones containing the gene. The locus corresponding to the \textit{uspC}, \textit{otsA}, and \textit{otsB} genes was most highly enriched at the end of the ethanol production selection.
Figure 4-3 Comparison of genome-wide enrichment in the production selection approach presented here to previously performed ethanol tolerance selections in a wild-type host strain (Woodruff et al., in press) using a similar SCALEs strategy. Previous tolerance selections were performed with 15 and 30 g/L of exogenous ethanol. All gene fitness values were calculated as described in this work...

Figure 4-4 Genotypes and ethanol tolerance of selected colonies. (A) Ten colonies were picked randomly from the plated selection culture at the end of eighth batch, and their genomic fragments contained on the plasmid were sequenced as indicated here. Five of the ten colonies contained fragments in the uspC-otsA-otsB genomic region, and these fragments are depicted in the right column. Partial genes are denoted (*), and clones C and F contain identical genomic fragments (**). (B) Selected colonies were grown in 15 g/L and 25 g/L of ethanol for 24 hr in 15 ml screw-cap tubes and otherwise identical to the selection environment. Only clones containing the fragments of the uspC-otsA-otsB genomic region show improved ethanol tolerance. Error bars represent the standard deviation from three biological replicates. P-values for comparison to the parent were calculated by a Student’s t-test (* p<0.05 in 15 g/L, ** p<0.1 in 25 g/L).

Figure 4-5 Absolute values of growth for selected clones grown in 15 and 25 g/L ethanol (corresponding relative values reported in Fig. 3B).

Figure 4-6 Ethanol production of selected clones. Batch fermentations in shake flasks were performed identical to the selection environment except that no exogenous ethanol was added initially. A sample was taken from each fermentation every 24 hr, and the cell density and ethanol concentrations were measured for all samples. Growth during the fermentation relative to the parent strain for (A) the clones containing the uspC-otsA-otsB genomic region and (B) all other selected clones is plotted. Ethanol production during each fermentation relative to the parent strain for (C) the clones containing the uspC-otsA-otsB genomic region and (D) all other selected clones is plotted. The average values for the empty vector strain are in parentheses next to the axis. Error bars represent the standard deviation of three biological replicate fermentations.

Figure 4-7 Ethanol tolerance and ethanol production of constructed clones to combinatorially test genes in the uspC-otsA-otsB genomic region. (A) Growth (OD600) in the presence of 25 g/L of ethanol (and otherwise production conditions) after 24 hr was measured for biological triplicates and compared to the empty vector (* p<0.05). Batch fermentations in shake flasks were performed and the (B) cell growth (OD600) and (C) ethanol produced relative to the empty vector parent strain are plotted, and the average values for the empty vector strain are in parentheses next to the axis. Error bars represent the standard deviation of three biological replicate experiments.
Figure 5-1 The trait-specific regulon engineering by assemblies of parts approach. (i) Data from a quantitative, genome-wide search for the desired trait, in this case a genome-wide selection for genotypes related to ethanol production in *E. coli* (Woodruff et al, submitted), (ii) is analyzed for trait-specific regulatory targets, for example by mapping the fitness onto the regulatory network. (iii) The identified trait-specific regulator is engineered by designing libraries which introduce genetic diversity into parts of the regulator which focus the genetic search space to mutations likely to alter cellular expression related to the trait of interest, for example by targeting the ribosomal binding site or the annotated DNA binding site with complete library coverage. (iv) Libraries of the engineered regulator are constructed by assembling the designed parts, subjected to competitive growth in a selective environment (or screened for the desired trait), and the fitness of the introduced mutations are quantified by high-throughput sequencing. The regulator can be recursively engineered for further strain improvements by designing successive libraries of parts based on the fitness landscape of the tested parts libraries.

Figure 5-2 Identification of regulatory targets related to ethanol production in *E. coli*. (A) The gene fitness values (peak height) from our previous genome-wide selection for ethanol production (Woodruff et al, submitted) for all genes related to transcription (COG K) are plotted at their corresponding genomic location for the fourth batch (blue inner circle) and final eighth batch (red outer circle) of selection. (B) The 500 genes with the greatest fitness at the end of the previous selection were mapped onto the regulatory network according to their regulators in order to identify global regulators that regulate expression of these genes related to ethanol production. The regulatory networks of the global regulators (C) *rpoH* (σ\(^32\)), (D) *rpoE* (σ\(^24\)), (E) *rpoS* (σ\(^38\)), and (F) *rpoD* (σ\(^70\)) are shown for reference with the number of genes regulated for each in parenthesis. Regulatory network diagrams were constructed using the EcoCyc regulatory overview (Keseler et al., 2011).

Figure 5-3 Design of the libraries of regulatory parts. (A) Genetic diversity was introduced into the regulator’s ribosomal binding site (RBS) and DNA binding site using custom oligos constructed with the desired mixed bases. The degenerate RBS design is identical to that used by MAGE (Wang et al., 2009), which introduces the degenerate RBS sequence (D=A,G,T; R=A,G) at the -4 to -14 positions upstream of the start codon. The 20 amino acid helix-turn-helix motif comprising the annotated DNA binding site of each targeted regulator was mutagenized by NNK (N=A,T,G,C; K=G,T) site saturated mutagenesis with each of the 10 custom oligos each containing two mutagenized codons the maximally distanced 9 residues apart. (B) Three libraries for each regulator were constructed to completely search the genotypic space of the RBS and DNA binding site parts as well as investigate some of the interactions between these parts with a combinatorial library that searches a small fraction of the combinatorial variants.
Chapter 1 Engineering genomes in multiplex


1.1 Abstract

Efficiently engineering robust complex traits is a key challenge facing metabolic engineering efforts to synthesize valuable products in vivo. Recent advances in genome engineering confront this barrier and significantly enhance the ability to map functional changes targeted throughout the genome and combinatorially optimize complex (multigenic) traits using multiplex recombineering. We describe a framework for efficiently searching genome-wide combinatorial space to optimize complex traits and highlight recent advances in genome engineering that enable this approach.

1.2 Introduction

The rapidly expanding collection of synthetic, molecular, and genome biology tools has enabled impressive achievements in rewiring cellular machinery to produce a variety of valuable chemicals, such as drugs (Ajikumar et al., 2010; Pfeifer et al., 2001; Ro et al., 2006; Wang et al., 2009), bioplastics (Lutke-Eversloh et al., 2002; Taguchi et al., 2008), platform chemicals (Atsumi et al., 2009; Dueber et al., 2009), and biofuels (Alper et al., 2006; Atsumi et al., 2008; Schirmer et al., 2010; Steen et al., 2010). As a result, substantial progress has been made in engineering an array of industrially relevant traits, ranging from high pathway flux and yields to product, feedstock, and process tolerance (temperature, pH, salinity, etc.) (Patnaik, 2008). Nevertheless, the task of engineering multiple complex traits into a host genome remains challenging and requires a detailed understanding of host biology. Directed evolution approaches have been valuable for addressing these challenges at the level of individual proteins (Turner,
2009). However, methods for rapidly introducing, screening, and recovering mutations have thus far been too slow or expensive to be applied comprehensively to complete pathways or genomes.

Recent advances in several technologies have now overcome this barrier. Improved DNA synthesis has dramatically advanced synthetic biology construction efforts and now permit rapid synthesis of dozens of genes in parallel (Kosuri et al., 2010). New homology-based ligation methods enable efficient pasting of these genes together into complete pathways and genomes (Gibson, 2009; Gibson et al., 2010). This technology for gene/genome construction is also being used to more rapidly identify strategies for rewiring genomes (Wang et al., 2009; Warner et al., 2010b). Recombination-mediated genetic engineering, known as recombineering, efficiently modifies host genomes using linear DNA and thus permits direct incorporation of synthetic DNA into targeted cell chromosomes (Datta et al., 2008; Ellis et al., 2001b; Murphy, 1998; Yu et al., 2000; Zhang et al., 1998). Coupled with cheap DNA synthesis, recombineering can produce thousands of genetic modifications in a single pool. In this review, we highlight the new approaches to genome engineering that take advantage of these advances, with a specific focus on the use of multiplex recombineering to simultaneously modify multiple genetic targets. We present a framework for optimizing complex genome-wide traits with minimal prior knowledge of genetic targets and discuss the recent advances in the tools necessary for efficient multiplex genome engineering.

1.3 A genome engineering algorithm

A fundamental barrier in strain engineering on the genome-scale is engineering complex traits which rely upon a combination of coordinated genetic changes. The difficulty arises from our ability to search comprehensively for optimal genetic combinations due to the finite throughput of our most effective selection and screening strategies. For example, if we test ten
expression levels for each gene to combinatorially optimize production of a given product, we could at best select or screen among the components of a pathway of up to fifteen genes, given a maximum laboratory throughput of approximately $10^{15}$ cells. However, combinatorial optimization of this nature becomes an impossible task on the genome scale. For even the smallest known genome, *Mycoplasma genitalium* with 485 protein-coding genes, there are now $10^{485}$ possible combinations; this number far exceeds the $10^{80}$ total number of atoms in the universe.

Similar to this genome-scale challenge, narrowing search space on the protein-scale in order to create functionally rich libraries within the limits of library size and laboratory throughput constraints has been at the forefront of directed protein evolution (Fox and Huisman, 2008; Romero and Arnold, 2009). Most single-nucleotide polymorphisms result in silent or detrimental effects on protein function. Empirical studies have shown that about 0.01-1% of total mutations result in a positive effect on protein function (Aharoni et al., 2005; Bloom et al., 2006; Guo et al., 2004; Rennell et al., 1991). Similarly, an evolved protein’s uphill climb on the fitness landscape is not a single path. There are a variety of solutions or paths to increased fitness, and certain paths may lead to local fitness maxima that require a descent in order to reach global maxima (Tracewell and Arnold, 2009). Pathway engineering studies confront similar search challenges. Single-gene perturbations are often not simply additive but instead require combinatorial approaches for optimization (Ajikumar et al., 2010; Liu et al., 2010). In the absence of a clear understanding of how different residue modifications might combine to improve function or how different genes in a pathway might combine to improve performance, directed protein and/or pathway evolution strategies have remained throughput limited. As we move engineering efforts to the genome scale (Gibson et al., 2010; Wang et al., 2009), these
limitations are exacerbated by the large increase in the number of genes that may confer functions relevant to the trait of interest.

To overcome this fundamental barrier in genome engineering, much can be learned from directed protein evolution algorithms developed over the last two decades (Arkin and Youvan, 1992; Bloom et al., 2006; Fox and Huisman, 2008; Lutz, 2010; Voigt et al., 2001). An analogous genome engineering algorithm is composed of three key steps (Figure 1-1). The first step is to expand and then directionally narrow the size of the genotypic search space (set of all possible genotypes) using synthetic biology based search technologies (Brynildsen and Liao, 2009; Goodarzi et al., 2010; Lynch et al., 2007; Warner et al., 2010b).

Such technologies produce a quantitative map of genetic modifications onto a trait of interest. Second, the relative relevance of these identified modifications is determined, either directly through the results of genome search technologies or through more complex calculations that consider a broader set of data. Finally, combinatorial optimization is carried out for the most relevant gene targets identified. This genome engineering algorithm narrows the vast search space of the genome from broad genome-wide libraries to trait-specific, functionally rich combinatorial libraries, largely paralleling the advancements in protein engineering strategy. Next we describe several recent advances that have enabled this directed approach to genome-engineering.
Figure 1-1 A genome engineering algorithm comprised of these three steps allows for an efficient search of the relevant combinatorial genotypic space for a trait of interest. (A) Genetic diversity on the genome-scale is constructed by targeting functional modifications to loci throughout the genome. A screen or selection for the desired trait is used to narrow the genetic diversity and enrich for individual genetic changes. (B) Relevance of the genetic changes is ranked based on results of the screen/selection and additional relevance metrics. Ranking and further narrowing the number of genes targeted directs the combinatorial search to comprehensively search the relevant genotypic space. (C) Genetic diversity targeted to fine-tune the most relevant genetic changes in parallel is introduced. A screen or selection for the desired trait is used to narrow the genetic diversity by enriching for the combinatorially-optimized, complex (multigenic) trait of interest.
1.4 Genome search technologies

Traditional strain adaptation methods rely upon recursive rounds of random mutagenesis and natural selection. Throughout this process the strain slowly acquires mutations, some of which beneficially impact the trait of interest. This approach is laborious and, as discussed above, beneficial mutations are infrequent and difficult to identify. A major limitation to this approach is that mutations are created randomly throughout the genome. Therefore, the vast majority of mutants in the resultant population are not improved. In response to this limitation, a range of genome-scale strategies have been developed for generating functional genetic diversity with a greater likelihood of beneficial mutations. Plasmid-based genomic overexpression libraries and insertional mutagenesis methods have been commonly used to successfully search the genome for individual genetic modifications that impact a given phenotype (Borden and Papoutsakis, 2007; Goodarzi et al., 2010; Lynch et al., 2007; Warnecke et al., 2008). The search space of these genome-scale methods, however, is limited by the library sizes that can be constructed and stably maintained within a host. Moreover, while recovery of the relevant mutations is vastly improved compared to prior approaches, challenges remain with respect to the resolution at which such mutations can be identified. To circumvent this limitation and allow for systematic screening or selection of mutants for all possible genes, collections of nonessential knockouts and overexpression of all open reading frames (ORFs) have been constructed for S. cerevisiae (containing molecular barcodes) and for E. coli (Baba et al., 2006; Giaever et al., 2002; Ho et al., 2009; Kitagawa et al., 2005; Yamamoto et al., 2009). These methods were recently reviewed and will not be detailed here (Santos and Stephanopoulos, 2008; Warner et al., 2009).
Multiplex DNA synthesis and recombineering technologies have enabled further improvements in efforts that generate libraries with increased levels of relevant mutations. Our group developed the trackable multiplex recombineering (TRMR, “tremor”) approach. Using TRMR we rapidly constructed and concurrently mapped promoter and ribosomal binding site (RBS) modifications for all genes in the genome (Warner et al., 2010b). Moreover, the TRMR approach employed a molecular barcoding strategy (Winzeler et al., 1999) to enable quantitative tracking of each genetically modified strain within a population of approximately 8,000 unique strains, each of which was engineered to increase or decrease expression of a particular gene in the \textit{E. coli} genome. While prior approaches have benefited from advances in genomics, the TRMR approach was primarily enabled by advances in synthetic biology and homologous recombination techniques.

Advances in parallel synthesis of DNA on microarrays make it possible to synthesize up to 55,000 unique 200-mer oligonucleotides (oligos) on a single microarray (Agilent) (LeProust et al., 2010). Microarrays with shorter oligos can be synthesized with more features, such as up to 4.2 million 75-mers (Roche Nimblegen) and 10 million 25-mers (Affymetrix). These oligo synthesis technologies allow for genome-wide functional libraries to be easily created using multiplex recombineering. Multiplex lambda Red recombineering was used to chromosomally insert the constructed oligos, containing the targeting homology arms, functional cassette and barcode tag, for each strain. Future efforts to engineer genomes using multiplex recombineering will be able to more easily construct large libraries due to the recent optimization and advancements in recombineering methods. Recombination frequencies of up to 50\% recombinant daughter cells, the theoretical maximum for recombination, have been achieved for single-stranded DNA (ssDNA) oligos and certain loci (Sawitzke et al., 2011). Some of the
modifications to improve recombination efficiency include: (1) using single-stranded DNA (ssDNA) oligos that target the lagging-strand (Ellis et al., 2001a; Mosberg et al., 2010), (2) using oligos of the optimal 70-90 bp length (Sawitzke et al., 2011; Wang et al., 2009), (3) adding phosphorothioate linkages to the four bases at the 5' terminus of the oligo (Mosberg et al., 2010), (4) designing oligos to avoid the methyl-directed mismatch repair system (Sawitzke et al., 2011), and (5) modifying the recombineering protocol, such as optimizing the culture growth, induction and mixing prior to electroporation (personal communication J. Sawitzke) (Sawitzke et al., 2011).

The use of molecular barcodes in \textit{in vivo} targeted mutagenesis genome search tools enables high-throughput monitoring of genome-wide modifications by either barcode microarray (Pierce et al., 2006) or reduced sequencing. As an alternative to microarray barcode quantification, the barcode analysis by sequencing, or “Bar-seq,” method has been developed to use next-generation sequencing to quantify barcodes present in a mixed population (Smith et al., 2009). Bar-seq was used to quantify several hundred thousand barcode-environment interactions simultaneously by carrying out a 96-plex experiment with 6,200 barcoded strains (Smith et al., 2010). Sets of molecular barcodes of up to 240,000 orthogonal 25-mer barcodes have been developed and facilitate creation of increased barcoded library size (Xu et al., 2009). Even if sequencing is used over microarrays to detect mutations or functional changes, using barcode tag-based approaches drastically reduces the sequencing necessary, as only the short barcode regions must be sequenced instead of the entire genome (Baird et al., 2008; MacLean et al., 2009; Smith et al., 2010).

Global approaches that generate combinatorial diversity at the transcription or translation level are also useful for searching genomes to identify genetic targets. These transcriptional and
translational engineering approaches have recently been reviewed in more detail than will be discussed here (Young and Alper, 2010). Global transcription machinery engineering (gTME) has been shown to be effective for engineering a variety of traits. Briefly, the gTME method constructs a mutant library of a global transcription regulator, such as a sigma factor or RNA polymerase subunit, in order to alter promoter specificity of the RNA polymerase complex and generate a diverse pool of transcriptomes (Alper et al., 2006; Alper and Stephanopoulos, 2007; Klein-Marcuschamer et al., 2009; Klein-Marcuschamer and Stephanopoulos, 2008; Klein-Marcuschamer and Stephanopoulos, 2010). Not all global transcription regulators are useful targets for improving all traits, because various regulator libraries improve some traits and not others. (Klein-Marcuschamer et al., 2009). Thus, these transcriptional engineering approaches require choosing trait-specific transcription regulators. In order to identify individual gene targets, transcriptional profiling identifies the different expression responsible for altering a phenotype. This differential expression in improved mutants has been shown to converge on a subset of genes’ expression changes through sequential rounds of gTME (Alper and Stephanopoulos, 2007). As an additional modification to the gTME method, mutating exogenous global regulators is also an effective strategy to engineer complex traits (Chen et al., 2011a).

A diverse set of tools now exists to effectively search the genome for trait-conferring genetic modifications, from overexpression or knockout libraries to in vivo targeted mutagenesis and global engineering approaches. These approaches to search the genome can be complementary. For example, TRMR can be used to identify a trait-specific regulator that could be a starting point for further optimization using a gTME approach in order to identify relevant gene expression changes. We believe that there is a class of in vivo targeted mutagenesis genome search tools that could be developed and accelerate the genome-wide mapping of functional
genotypes to phenotypes. With customized oligos, these tools could investigate more than two functional mutations per gene and a wide array of functions, such as increased, decreased, or a range of rates of transcription and/or translation. Additionally, targeted mutagenesis of the genome can be implemented recursively to build up multiple allelic changes and combinatorially engineer complex traits in vivo.

1.5 Assigning relevance

The second step in our genome engineering algorithm (Figure 1-1) is to rank the relevance of genetic targets that are likely to be beneficial through combinatorial engineering. Much like a web browser ranks the relevance of webpages related to an internet search query, it is necessary to assign relevance to gene targets across the genome based on a genome search for a given trait. This type of relevance assigning is not new to directed evolution, in fact it represents the basis of several powerful protein engineering algorithms (Estell and Wells, 1988; Fox et al., 2007; Voigt et al., 2001). Fox and colleagues demonstrated the power of such directed search strategies in their protein sequence activity relationships (ProSAR) mapping strategy (Fox et al., 2007). The authors mapped specific amino acid modifications onto protein activity and then, using statistical algorithms, identified a subset of residues to target exhaustive combinatorial searching. Similarly, Estell and Wells, mapped the effect of every residue modification at specific positions prior to combinatorial optimization (Estell and Wells, 1988). Likewise, Voigt et al. took a powerful yet distinct approach by first using computational design strategies to identify relevant residues to target (Voigt et al., 2001).

At the genome scale, much work is still required to determine effective relevance metrics. The genome searches described above utilize high-throughput screens and selections to map genetic modifications onto traits of interest. It has been shown, however, that selection design
has a large effect on the enrichment of various clones within a mixed-population selection culture (Gall et al., 2008a; Warnecke et al., 2008). To account for this variability, relevance ranking strategies might use fitness values obtained from multiple selection designs. These designs can be tailored selection strategies for the particular desired characteristics of the trait, such as increased growth rate or inhibitory concentration in a stressful environment or range of process conditions. Moreover, relevance metrics could include a range of data that goes beyond experimental selection or screening results, such as functional relatedness, pathway or other network co-occurrence, gene ontology patterns, and/or other strategies for inferring gene-to-gene relationships. This broader approach to relevance is similar in scope to combining the various rational, computational, and evolutionary protein engineering strategies described above. We expect that directed genome evolution approaches will find design guidelines for engineering complex traits. Some of these guidelines will be intuitive and some will be surprising as has been the case in directed protein engineering efforts (such as the finding that increased stability promotes evolvability (Bloom et al., 2006)).

1.6 Combinatorial optimization

The intricacy of biological networks obscures the predictability of genotype-phenotype correlations for complex traits and requires tuning the genome \textit{in vivo} to test combinations of genetic modifications via screens or selection. Hence, genome engineering algorithms require strategies to carry out combinatorial optimization of the expression of the most relevant genes (Figure 1-1C). As depicted in Figure 1-2, the number of possible combinations grows rapidly as the number of genes targeted for combinatorial optimization increases. Therefore, as the number of relevant genes targeted increases, it becomes increasingly difficult to test all possible combinations of genetic targets. With multiple expression states for each gene target additional
dimensions are added to the combinatorial space, and the number of combinations grows exponentially with the number of genetic targets. The ranking step described above is important for focusing the combinatorial genotypes to the most relevant search space.

Figure 1-2 Combinatorial search space. The combinatorial genotypic space (number of possible genotypes comprised of combinations of genetic targets) grows rapidly when additional loci are targeted as depicted by the combinatorial space for (A) 2 loci (B) 4 loci and (C) 6 loci, where each letter represents an individual locus. Figure is adapted from Wright’s original adaptive landscapes (Wright 1932).

Pathway engineering tools typically utilize systematic multivariate or combinatorial optimization of a few enzyme-encoding gene targets in order to maximize production of the desired product. Systematic multivariate optimization approaches for balancing fluxes within a pathway include synthetic protein scaffolds to localize and stoichiometrically balance enzymes (Dueber et al., 2009), a modular vector system to carry out gene dosage experiments (Anthony et al., 2009) and multivariate-modular pathway engineering which optimizes expression of modules or segments of the entire pathway (Ajikumar et al., 2010). Combinatorial pathway engineering tools generate libraries with varied expression levels of pathway constituents using methods such
as promoter engineering which could be used to combinatorially test a range of promoters strengths for each enzyme (Alper et al., 2005) and tunable intergenic regions (TIGRs) which generate variable mRNA secondary structures that alter the stability of individual transcripts encoding pathway components (Pfleger et al., 2006). These pathway engineering tools could be used to combinatorially optimize a few relevant genes identified. However, complex traits typically have complicated network interactions, and therefore, predicting a few genes which will have an additive effect on the phenotype is difficult.

There are also global approaches to combinatorially optimize complex traits, which include whole genome shuffling by protoplast fusion or conjugation. Genome shuffling allows for homologous recombination events among entire heterologous genomes within the genetically diverse population conferring the improved trait of interest. Genome shuffling has been used to engineer a variety of traits related to improving product yield, tolerance to inhibitors or substrate uptake in many different industrially relevant bacterial and yeast strains (as recently reviewed (Gong et al., 2009)). With these genome shuffling approaches, the relevance of the combinatorial mutation space is increased at the expense of the ability to identify the trait conferring mutations. Nonetheless, continued improvements genome sequencing will help to more rapidly identify these beneficial genetic changes and, as a result, we expect the use of genome shuffling methods to increase going forward.

Multiplex automated genome engineering (MAGE) uses parallel DNA synthesis and multiplex recombineering to combinatorially optimize the expression of a sub-set of relevant genes. The MAGE approach was used to improve lycopene production five-fold within three days by carrying out recursive allelic replacement with degenerate RBSs for 24 genes related to the DXP pathway (Wang et al., 2009). The MAGE approach utilized 90 bp ssDNA oligos that
target the lagging strand of target genes for λ Red β protein facilitated homologous recombination. Importantly, Wang and his colleagues developed an automated system to carry out recursive rounds of allelic replacement, which enabled the generation of a cumulative pool of roughly $10^9-10^{10}$ mutants. These were so-called rational mutants, where the specific location of all introduced mutations was known a priori. Their best lycopene-producing strain (5x improved) isolated after 35 cycles was identified by screening $10^5$ colonies (~0.01% of the total $~7 \times 10^8$ population size). Since the mutation locations were known, the authors were able to rapidly determine that this most improved strain contained modifications for 5 of the targeted genes. Using recursive and automated cycles of recombineering, MAGE enables the rapid construction of a vast combinatorial space of genetic variants.

1.7 Conclusions

New technologies in synthetic, molecular, and genome biology have enabled the extension of powerful directed protein evolution algorithms to the genome scale. These new technologies are beginning to find use in academic and industrial strain engineering efforts. However, the integration of such tools into a coherent framework is only now beginning. The recursive and combinatorial nature of MAGE makes it well suited for fine-tuning expression of relevant gene targets identified by an efficient genome search. The MAGE approach, however, requires prior knowledge of the gene subset to be targeted. Therefore, genome search technologies, such as TRMR, can enable MAGE-like combinatorial optimization by identifying the most relevant genetic modifications. The coupling of TRMR and MAGE would allow for the fine-tuning of the specific functions to generate a more relevant combinatorial search space, such as degeneracy designed for stronger or weaker RBSs based on the genome search findings. This coupling of new approaches represents a first example of the directed genome evolution
algorithm described above. However, it is important to note that many other technologies that could be generalized as search, relevance, and combinatorial optimization methods exist or are on the horizon. One particularly exciting future opportunity involves the integration of new synthetic biology circuits and biological parts with the approaches described herein (Carr and Church, 2009; Lu et al., 2009; Voigt, 2006). It is conceivable that such circuits could be used to expand the range of genetic modifications employed in search and optimization strategies as well as to enable the tracking of targeted cellular states \textit{in vivo}. In combination, these approaches represent the extension of rational design and directed evolution from the protein scale to the genome scale.

\textbf{Acknowledgements}

LBAW was supported by a Graduate Research Supplement from the US National Science Foundation CBET0449183.

\textbf{1.8 References}


Chapter 2 Genome-scale identification and characterization of ethanol tolerance genes in *Escherichia coli*

In press at *Metabolic Engineering* (November 2012).

2.1 Abstract

The identification of relevant gene targets for engineering a desired trait is a key step in combinatorial strain engineering. Here, we applied the multi-Scalar Analysis of Library Enrichments (SCALEs) approach to map ethanol tolerance onto 1,000,000 genomic-library clones in *Escherichia coli*. We assigned fitness scores to each of the ~4,300 genes in E. coli, and through follow-up confirmatory studies identified 9 novel genetic targets (12 genes total) that increase *E. coli* ethanol tolerance (up to 6-fold improved growth). These genetic targets are involved in processes related to cell membrane composition, translation, serine biosynthesis, and transcription regulation. Transcriptional profiling of the ethanol stress response in 5 of these ethanol-tolerant clones revealed a total of 700 genes with significantly altered expression (mapped to 615 significantly enriched gene ontology terms) across all five clones, with similar overall changes in global gene expression between two clone clusters. All ethanol-tolerant clones analyzed shared 6% of the overexpressed genes and showed enrichment for transcription regulation-related GO terms. iTRAQ-based proteomic analysis of ethanol-tolerant strains identified upregulation of proteins related to ROS mitigation, fatty acid biosynthesis, and vitamin biosynthesis as compared to the parent strain’s ethanol response. The approach we outline here will be useful for engineering a variety of other traits and further improvements in alcohol tolerance.
2.2 Introduction

Identifying relevant gene targets is a critical step in the process of engineering and optimizing a strain for a desired trait (Alper et al., 2006; Borden and Papoutsakis, 2007; Brynildsen and Liao, 2009; Tong et al., 1991; Withers et al., 2007; for a recent review see Woodruff and Gill, 2011). Adaptation strategies, which rely upon natural random mutagenesis, have been proven an effective means for obtaining an improved strain for a wide variety of traits (Cooper and Lenski, 2000; Herring et al., 2006; Minty et al., 2011; Smith and Liao, 2011; Yomano et al., 1998). Nevertheless, even with relatively affordable genome sequencing, uncovering the beneficial mutations remains laborious since the probability of deleterious or neutral mutations is much greater than that of beneficial mutations (Barrick et al., 2009; Kimura, 1983). In comparison to traditional adaptation, newer genome search strategies enable the more rapid mapping of beneficial genetic changes. These strategies introduce diversity through various molecular approaches, including the use of plasmid-based genomic libraries (Gill et al., 2002; Kitagawa et al., 2005; Lynch et al., 2007), transposon libraries (Baba et al., 2006; Badarinarayana et al., 2001), or multiplex recombineering-enabled libraries (Warner et al., 2010a). Advanced sequencing methods or microarray analysis are then applied to track such libraries throughout various selections or screens (Amini et al., 2009; Bonomo et al., 2008; Gall et al., 2008b; Goodarzi et al., 2010; Warnecke et al., 2008). The multi-Scalar Analysis of Library Enrichments (SCALEs) genome search strategy (Lynch et al., 2007) has been used in previous studies to elucidate the influence of selection design on the selected phenotypes, demonstrate high reproducibility of selections, and map the genome for genotypes conferring tolerance to a variety of inhibitory compounds (Bonomo et al., 2008; Gall et al., 2008b; Sandoval et al., 2011; Singh et al., 2009; Spindler et al., 2011; Warnecke et al., 2010; Warnecke et al.,
Here we applied the SCALEs approach to identify genetic elements for which overexpression improves growth of \textit{E. coli} under ethanol stress.

We chose to study ethanol tolerance in \textit{E. coli} not only because of the significant commercial importance of this trait but also to serve as a model trait since ethanol is among the best-studied for stress effects and tolerance mechanisms of the industrially-relevant compounds. In general, it has been shown that the toxicity of an alcohol strongly correlates to its hydrophobicity, and this correlation has been ascribed to disruption of the cell membrane by the alcohol molecules partitioning directly in the lipid bilayer and hydrogen bonding to polar membrane constituents (Ingram, 1990; Ingram and Buttke, 1984). In \textit{E. coli}, ethanol exposure is known to inhibit biosynthesis of constituents of peptidoglycan, fatty acids, and lipids (Buttke and Ingram, 1978; Clark and Beard, 1979; Ingram, 1977). More broadly, ethanol has been shown to compromise membrane integrity and lead to leakage of intracellular ions and molecules, dissipation of the proton motive force, and inhibition of macromolecular biosynthesis (Fried and Novick, 1973; Ingram, 1990; Ingram and Buttke, 1984; Kobayashi et al., 2007). Additionally, in yeast, superoxide dismutase has been found to aid in the mitigation of toxic oxygen radical species produced at elevated ethanol concentrations (Alexandre et al., 1994; Costa et al., 1993; Salgueiro et al., 1988).

Due to the probable complexity of the mechanism of ethanol inhibition, engineering of ethanol tolerance has remained a major challenge. Even so, a number of strain engineering approaches have engineered improved ethanol tolerance. Long-term adaptation of an ethanol production strain by serially culturing the strain in ethanol over a few months identified an improved strain that grew substantially better in ethanol and produced more ethanol (Yomano et al., 1998). This improved strain was characterized by transcriptional profiling, resulting in the
identification of approximately 200 genes with altered expression (Gonzalez et al., 2003). As with almost all transcriptional profiling efforts, discriminating between the genes whose altered expression confers tolerance and those whose expression changes as a result of tolerance mutations or other mutations in the adapted strain was a challenge. A few genome-wide studies on ethanol tolerance have identified genes for which altered expression or mutation resulted in improved tolerance, and these studies have shown that rewired regulatory networks are a valuable approach for engineering complex phenotypes eliciting ethanol tolerance (Alper et al., 2006; Alper and Stephanopoulos, 2007; Goodarzi et al., 2010).

Here, we have extended these prior studies by mapping the effect of increased expression for all genes in the genome onto ethanol tolerance using our SCALEs method (a total of 1,000,000 overlapping genomic-library clones are included in our mapping efforts). We identified many novel trait-conferring genetic targets that improve growth in ethanol that are not regulatory proteins and are involved in a variety of pathways and cellular processes. In an effort to better understand the overlap in tolerance phenotypes conferred by these genes, we compared genome-wide changes in gene expression and protein expression for a subset of confirmed ethanol-tolerant clones.

2.3 Materials and methods

2.3.1 Strains, media, and reagents

Genomic DNA was extracted from *E. coli* K12 (ATCC #29425) cultured overnight in Luria-Bertani (LB) medium. Genomic libraries were prepared in the pSMART-LCKan vector (Lucigen). *E. coli* BW25113 Δ*recA::FRT* was used as the host strain for genomic libraries and constructed clones; BW25113 Δ*recA::Kan* was obtained from the Keio collection (Baba et al.,
2006), and the Kan cassette was removed following the previously described method (Datsenko and Wanner, 2000). Selections and growth testing were carried out in MOPS minimal medium with 2 g/L dextrose as the sole carbon source (Neidhardt et al., 1974). All cultures were incubated at 37°C. Molecular biology grade ethanol was used for ethanol dosing (Sigma-Aldrich #E7148). Primer synthesis and sequencing was performed by Operon. Kanamycin, 30 μg/ml, was added to all cultures containing the pSMART-LCKan vector.

2.3.2 Genomic library preparation and selections

Genomic libraries were previously prepared by Warnecke et al. in MACH1-T1 (Invitrogen) and were stored at -80°C in glycerol (Warnecke et al., 2008). In this study, we used genomic libraries containing 1, 2, 4, 8 kb genomic DNA (gDNA) fragments. Plasmid DNA of each of the 4 genomic libraries was purified by the QIAprep Spin Miniprep Kit (Qiagen). A BW25113 ΔrecA::FRT culture was made electrocompetent by a water and glycerol wash protocol (Sambrook and Russell, 2001) and then transformed with each of the plasmid libraries by electroporation. A yield of >10⁶ transformants was obtained for each library, which ensures a 99.9%+ probability of the entire genome being represented at approximately 125 bp intervals. Transformations were recovered in SOC media for 1 hour before being added to 100 ml MOPS media with kanamycin and grown to an optical density at 600nm (OD₆₀₀) of 0.4 in a 250 ml shake flask.

Selections were inoculated to an OD of 0.04 in MOPS medium with ethanol dosed to a final concentration of 15 g/L or 30 g/L. The 50 ml selection cultures were grown in 125 ml shake flasks with rubber septa to prevent evaporation and a 22 gauge needle to vent. The selections were serially transferred every 12 hours with the exception of the second batch of the 30 g/L selection which required 60 hours to show growth of at least one doubling (OD₆₀₀>0.08). We
have repeatedly observed a long lag phase for the second batch of growth for a variety of libraries when grown under high ethanol stress. Samples were plated from the selection starter culture and the final batch of each selection for microarray analysis.

2.3.3 Quantification of selection enrichment by microarray analysis

Samples for microarray analysis were prepared by plating approximately $10^5$ clones of the selection culture on LB plates with kanamycin. Plates were incubated at 37°C for 12 hours before gently scraping and pelleting the cells. Plasmid DNA from the cell pellets was extracted using the HiSpeed Plasmid Midi Kit (Qiagen) and then concentrated by ethanol precipitation. Further preparation of the plasmid DNA samples for the microarrays, microarray analysis and analysis using the SCALEs software was carried out as described previously (Warnecke et al., 2008). This software was used to quantify clonal enrichment and clone fitness values ($W_{\text{clone}}$).

2.3.4 Gene fitness determination

The gene fitness ($W_{\text{gene}}$) values for all genes in the EcoCyc database (Keseler et al., 2011) were calculated from the fitness of all clones determined by the SCALEs software. A gene’s fitness was calculated by summing the clonal fitness for all clone’s containing at least 75% of the length of the gene. In general, we do not expect that expressing truncated genes resulted in an effect in a clone’s ethanol tolerance phenotype (though we cannot rule out this possibility), and for this reason, we cloned the whole ORF when testing identified genes for ethanol tolerance. A 75% threshold was used because the Affymetrix gene-chips employed here do not contain probes for the entire gene, and as a result the SCALEs algorithm is limited to $\geq125$ nt resolution of the genome. Thresholds below 75% did not substantially alter the gene fitness values or highest fitness genes, yet a 100% threshold eliminated 4 of the highest fitness genes and 3 of these genes were later confirmed to confer ethanol tolerance.
2.3.5 *Clone construction*

Genomic regions were amplified with 5'-phosphorylated primers designed for genes of interest (primer sequences in Table 2-1) using Platinum *Pfx* DNA polymerase (Invitrogen) according to the manufacturer’s recommendations. PCR products were purified by gel electrophoresis and a QIAquick Gel Extraction kit (Qiagen). Purified PCR products were ligated to the pSMART-LCKan vector using a CloneSMART kit (Lucigen). Ligations were incubated at 4°C overnight followed by 2 hours at 25°C. Ligation reactions were ethanol precipitated and transformed into BW25113 Δ*recA*::FRT that was made electrocompetent by a water and glycerol wash protocol (Sambrook and Russell, 2001). After a 1 hour recovery, transformations were plated on LB plates with kanamycin. Clone sequences were verified by sequencing plasmid DNA extracted using a QIAprep Spin Miniprep Kit (Qiagen) with primers SL1 5'-CAGTCCAGTTACGCTGGAGTC-3' and SR2 5'-GGTCAGGTATGATTTAAATGGTCAGT-3', and insert sizes were verified by gel electrophoresis to check for chimeras.

<table>
<thead>
<tr>
<th>Clone</th>
<th>Gene(s)</th>
<th>Forward primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>lpcA</td>
<td>5'-AAAGCCCCCTTACTTGTAGGAGGTCTGA-3'</td>
<td>5'-TCGCATCAGGCATCACGCCACAAT-3'</td>
</tr>
<tr>
<td>2</td>
<td>tiiS</td>
<td>5'-TGCGTGGTGAAGCTATGAGCAGT-3'</td>
<td>5'-TGCGTCTTATCCGGCTTTCTACTTAAAC-3'</td>
</tr>
<tr>
<td>3</td>
<td>yhfT, yhfU</td>
<td>5'-TTAAGCCGGAACAGACAGGCACAG-3'</td>
<td>5'-GGTGCGTGAACCCGCTATGCTGTC-3'</td>
</tr>
<tr>
<td>4</td>
<td>fadE</td>
<td>5'-TTAGCGGCTTCAACTTGGTCACAC-3'</td>
<td>5'-TACCCGATACCACAAAAGCGGAAAC-3'</td>
</tr>
<tr>
<td>5</td>
<td>serA</td>
<td>5'-CGGTGTGGAGAAGGGATAAA-3'</td>
<td>5'-AAAATCATTTGCTTCCGGTGC-3'</td>
</tr>
<tr>
<td>6</td>
<td>zur</td>
<td>5'-GTAGCCCTTACTAAGCCTTTT-3'</td>
<td>5'-TGGATTACACCCGGCTTTATCTTCA-3'</td>
</tr>
<tr>
<td>7</td>
<td>armC, armB</td>
<td>5'-TCTGCGATAATCGTACGTTTGTG-3'</td>
<td>5'-ATCGTGTTAGCCAAAAACGACGGT-3'</td>
</tr>
<tr>
<td>8</td>
<td>yicE</td>
<td>5'-CTGCTACACTATCCGCGAGTTG-3'</td>
<td>5'-ACCCGCCAGTGAAATTTCTTTC-3'</td>
</tr>
<tr>
<td>9</td>
<td>yaeQ, yaeJ</td>
<td>5'-TTAACCCTGAGTGAATCTACTAAGG-3'</td>
<td>5'-TTATTCCGACCAGCTGACACTTTG-3'</td>
</tr>
<tr>
<td>10</td>
<td>yafK</td>
<td>5'-TTATTTGCCCTGGGGAGCCTGTA-3'</td>
<td>5'-AGCCCGCTCAGATGAAATTTTGA-3'</td>
</tr>
<tr>
<td>11</td>
<td>yafJ</td>
<td>5'-TGCGGCGATCAGTATGAGCAGAAAT-3'</td>
<td>5'-TAACTACACCGTCCCGAGG-3'</td>
</tr>
<tr>
<td>12</td>
<td>trmA</td>
<td>5'-TTGCTAGCTACGTTACGAGTG-3'</td>
<td>5'-AGTACTGTCAGCTCAGTAC-3'</td>
</tr>
<tr>
<td>13</td>
<td>yijD, fabR</td>
<td>5'-CGCTGTACGTAAGAAGACCCGGCA-3'</td>
<td>5'-ACTGACCGGCAAGTAAACCCCATG-3'</td>
</tr>
</tbody>
</table>

*Table 2-1* Primers used to amplify genomic regions containing genes of interest.
2.3.6 Growth testing and ANOVA

Cultures were inoculated from freezer stocks into 5 ml of MOPS media with kanamycin in a 15 ml screw-cap tube. After 24 hr of growth, the overnight cultures were used to inoculate 5 ml starter cultures to an OD$_{600}$=0.1. The starter cultures were grown to mid-exponential phase (OD$_{600}$=0.4) and diluted as necessary. Growth of the cultures in ethanol was tested in flat-bottom 96-well microtiter plates (Corning #3370) with 100 μl total volume and inoculated at OD$_{600}$=0.04. To each well, 10 μl inoculum of starter culture diluted to OD$_{600}$=0.4 and 90 μl of MOPS media with kanamycin and ethanol dosed for a final concentration of 25 g/L ethanol. Each clone’s growth was tested with 8 replicate wells randomized on each plate from the starter culture and tested in triplicate on 3 plates each ran on separate days (n=24 per clone). The 14 clones were split randomly into two subsets for each replicate plate, but the empty vector control growth was tested on all of the six plates (n=48 for control strain). The plates were sealed with packing tape to prevent ethanol loss and incubated at 37°C with shaking in a PowerWave XS microplate spectrophotometer (Bio-tek). A two-way ANOVA was performed using Minitab to compare each clone’s growth to the control empty vector strain for 8 replicates on each of the clone’s triplicate plates. Growth was calculated for each well as the initial OD$_{600}$ subtracted from the OD$_{600}$ after 16 hours.

2.3.7 Transcriptional profiling

The ethanol stressed samples for transcriptional profiling were prepared nearly identically to the growth testing but with a larger volume. Overnight cultures and starter cultures were prepared identical as above (section 2.6). Starter cultures were diluted to OD$_{600}$=0.4 and used to inoculate 20ml cultures to an OD$_{600}$=0.04 in MOPS media with kanamycin and ethanol dosed to a final concentration of 25 g/L. These ethanol-stressed cultures were grown in 50 ml screw-cap
tubes at 37°C with 225 rpm shaking for 5 hours. The cells were then pelleted by centrifugation at 5,000 rpm for 6 min and resuspended in 400 μl of media supernatant. The resuspended cell pellets were mixed into microcentrifuge tubes with 800 μl RNAprotect Bacteria Reagent (Qiagen) and incubated at room temperature for 5 minutes before centrifugation for 10 min at 13,000 rpm. Supernatant was decanted, and cell pellets were frozen at -80°C. Cell pellets for each clone and the control were prepared in duplicate. RNA was extracted from cell pellets using an RNeasy Mini kit (Qiagen) with RNase-Free DNase (Qiagen) digestion. Transcriptomes were quantified on E. coli Antisense Genome Arrays (Affymetrix). Synthesis of cDNA, cDNA fragmentation, and terminal labeling was performed according to the manufacturer’s recommended protocol with the listed reagents (Affymetrix GeneChip Expression Analysis manual Chapter 3). Target DNA was hybridized to the microarrays, washed, stained, and scanned according to the manufacturer’s protocol at the University of Colorado Microarray Facility using a GeneChip Hybridization oven, GeneChip Fluidics Station, GeneArray scanner, and GeneChip Command Console Software (Affymetrix). The Expression Console Software (Affymetrix) was used to perform quantile normalization and background correction for all arrays with the Robust Multichip Analysis (RMA) workflow. Probe signals were converted to concentrations based on the correlation constructed from the control probe signals and respective concentrations. For all five ethanol-tolerant clones the largest fold-change in expression compared to the parent strain was for the gene or genes contained on the constructed plasmid, which ranges from a 6- to 30-fold increase in concentration.

2.3.8 iTRAQ proteomics

Biological duplicate cell pellets for clones and the empty vector control were prepared and grown in ethanol identical to the cultures used for transcriptional profiling (section 2.7). After 5
hours of growth in 25 g/L ethanol, cultures were centrifuged for 10 min at 5,000 rpm. Cell pellets were resuspended in 100 μl PBS (pH 7.4), frozen in liquid nitrogen, and stored at -80°C until shipment to the University of Sheffield on dry ice.

Intracellular soluble proteins were extracted and quantified using methods previously described (Pandhal et al., 2011). 100 μg of protein for each sample was digested and labeled with 8-plex iTRAQ isobaric reagents following the manufacturer’s instructions (AB SCIEX). Labeled peptides were combined and subsequently fractionated using HILIC chromatography using parameters previously defined (Ow et al., 2011). The resulting 35 fractions were dried in vacuo prior to LC-MS/MS analysis. Dried peptide fractions were dissolved in 17 μL of Switchos buffer (3% ACN and 0.1% formic acid), and a 5 μL aliquot was injected into an Ultimate 3000 nanoflow reverse phase liquid chromatography system (Dionex) coupled to a QStar XL Qq-TOF nano-ESI-MS/MS system (AB SCIEX). Separations were achieved with a PepMap 75 μm C-18 fused silica column (Dionex) performed at a constant flow rate of 300 nL/min. Liquid chromatography conditions used were identical to those described previously (Pandhal et al., 2011). The mass spectrometer was set to acquire in the positive ion and data-dependent acquisition mode, with the two most intense ions selected for CID fragmentation. Other parameters for mass spectrometry were set as described elsewhere (Pandhal et al., 2011). Protein identifications were carried out using Phenyx v2.6 (GeneBio) and quantitative analysis was performed using in-house developed programs as previously described (Pandhal et al., 2011). In brief, protein quantifications were obtained from computing median normalized geometric means of the identified peptide iTRAQ reporter ion intensities. Protein identifications were validated via decoy-target databases on a maximum of 1 % false discovery rate and a minimum of 2 identified peptides per protein. Significant changes in protein expressions were estimated
via a one-tailed $t$-test at a $p < 0.05$ (95% confidence interval) with Bonferroni test correction as detailed previously (Pham et al., 2010).

2.3.9 Clustering and gene ontology enrichment analysis

Hierarchical clustering was performed using GenePattern with average linkage and Pearson correlation of row and column normalized concentrations of all probes for each RMA normalized transcriptional profiling array (Reich et al., 2006). For gene ontology enrichment analysis, GOMiner was used with EcoCyc GO database and EcoCyc gene name identifiers (Zeeberg et al., 2003). Venn diagrams were generated using VennMaster version 0.37.5 (Kestler et al., 2008).

2.4 Results and discussion

2.4.1 Genome-wide selections for ethanol tolerance

We employed the SCALEs approach to map genetic regions conferring ethanol tolerance. The SCALEs approach utilizes multiple genomic libraries each comprised of genomic fragments of a specific size. These multiple genomic libraries are mixed together, exposed to a growth selection, and tracked at the individual clone level by DNA microarrays (Lynch et al., 2007). In this study we used genomic libraries with 1, 2, 4, and 8 kb genomic fragments of wild-type *E. coli* K12 and ensured that each library had greater than $3 \times 10^5$ transformants for 99.9% probability of the entire genome being represented at 125 bp intervals on average (Figure 2-1).
Figure 2-1 Design of ethanol tolerance selection strategy utilizing the SCALEs approach. Four plasmid libraries containing either 1, 2, 4, or 8 kb fragments of E. coli genomic DNA were constructed and in sufficient numbers to ensure 99.9% probability of entire genome representation in each library. Serial batch selections were carried out with 15 g/L or 30 g/L ethanol addition. Microarrays were used to quantify the population clones at the end of each selection as well as the starting library population.

We performed serial batch selections on these mixed genomic libraries at ethanol concentrations of either 15 g/L or 30 g/L for all batches, which corresponds to a moderate or high level of ethanol stress, respectively. Each selection was serially transferred every 12 hours or until the selection culture showed at least one doubling of growth. The selections with moderate and high ethanol stress were carried out for approximately 25 and 10 generations of growth, respectively. By choosing ethanol concentrations less than the minimum inhibitory concentration, we designed these selections to investigate increased growth in the presence of ethanol as opposed to survival (Gall et al., 2008b). Samples of the plasmid DNA in the initial mixed genomic libraries and the final enriched population from each selection were quantified by genome microarrays and the SCALEs software. The enrichment of each clone during the
selection was calculated as the clonal fitness, $W$, which is defined here as the concentration of the clone after selection divided by the concentration of the clone before selection.

162 clones in the 15 g/L ethanol selection and 272 clones in the 30 g/L ethanol selection were enriched ($W>1$) in each selection (Figure 2-2A, Figure 2-2B). This represents a reduction in initial library diversity ($4\times10^6$ transformants in mixed libraries) of greater than 99.99%, which suggests a strong overall selective pressure in both enrichments. The slightly greater reduction in overall diversity observed for the 15 g/L selection is likely to be largely dependent upon the 2.5-fold greater number of growth generations. However, the relative distribution of growth rates for clones in the population also plays a role in the reduction of diversity, and this distribution is dependent upon factors such as the selective pressure and the particular library. Interestingly, many genomic regions were highly enriched in both selections with moderate or high levels of ethanol stress with 40% of genomic regions (at 125 bp intervals) enriched in the 15 g/L selection also enriched in the 30 g/L ethanol selection. Some of these regions include the $tilS$–$yaeJ$, $fadE$–$yafK$, $serA$, $yicE$, $yijD$ loci. It should be noted, however, that there are substantial differences in the relative fitness values and rankings of the genomic regions enriched in each selection.
Figure 2-2 Enrichment of genomic library clones during ethanol selections. (a) Chromosomal enrichment and fitness of clones enriched during 15 g/L ethanol selection and 30 g/L ethanol selections. Each peak represents the natural log fitness, \(\ln(W)\), of a clone containing a genomic fragment corresponding to its location on the chromosomal map (genomic start position at twelve o’clock). Peak color denotes the size of the clone’s genomic fragment. (b) Histogram of the clonal fitness for all unique clones quantified in each selection.
2.4.2 Identification and confirmation of ethanol tolerance genes

We converted the clonal fitness values into gene fitness values in order to correlate the ethanol tolerance phenotype to known gene product functions and larger cellular processes (Figure 2-3A). Gene fitness values were calculated as the sum of the fitness values for all clones that contained at least 75% of the length of a given gene’s open reading frame (ORF). We do not expect that in our selected clones we were expressing truncated genes to affect ethanol tolerance. Rather, the threshold accounts for limitations in accurately determining the exact genomic fragment in a clone using the microarray analysis (e.g. genes that do not have probes targeting the entire length of the gene, the 125 nt genomic resolution of the SCALEs fitness mapping). Using this approach, we confirmed that both of our selections were indeed strong, with only 158 or 487 of the ~4300 genes in *E. coli* enriched \((W_{\text{gene}} > 1)\) in the 15 g/L and 30 g/L selection, respectively. The fitness values in each selection were non-normally distributed, with the top 10% of enriched genes exhibiting a broad range of fitness values much greater than 1 (top 10%: \(14 < W_{\text{gene}} < 140\) in 15 g/L selection and \(4 < W_{\text{gene}} < 124\) in 30 g/L selection). The enriched genes for each selection overlapped considerably for each selection, yet the relative rank of such genes changed considerably in each selection (i.e. *yafK, fade, yhfUT, serA*).
Figure 2-3 Gene fitness and ethanol tolerance of top fitness genes. (a) Fitness of each gene (calculated as the sum of clone fitness values for all clones containing at least 75% of the gene) in each selection. Highest fitness genes from selections are colored and labeled, and clones containing these genes with their native promoters were constructed. (b) Ethanol tolerance of constructed clones containing top genes was measured in a microtiter plate at 16 hours of growth. Error bars for each clone indicate standard deviation for 8 replicate wells on 3 separate plates (n=24). Empty vector control repeated on each plate (n=48). P-values were calculated by a 2-way ANOVA with factors accounting for clone and plate.
We next cloned 14 of these high-fitness genes (approximately evenly split between the two selections) along with their annotated or probable promoters onto the same plasmid vector used in the ethanol selections so that we could test the effects of these functional gene products on ethanol tolerance (Table 2-2). When constructing these clones, we included the native promoter and the entire ORF for these genes because we expect overexpression of the functional gene products to be the common route to phenotypic changes using genomic libraries. Some of these clones contain more than one gene if they are located in an operon. We tested the growth of each of these fourteen clones in 96-well plates with 25 g/L ethanol added to the media. An ethanol concentration of 25 g/L ethanol was chosen since this showed the maximal improvement in growth. The growth data was analyzed using a two-way analysis of variance. Nine of the fourteen clones showed improved growth (P<0.001) under ethanol stress as compared to the parent empty vector strain (Figure 2-3B), with up to a 5.9-fold improvement in the amount of growth for the lpcA clone. Three of the five clones that did not show improved ethanol tolerance (yafJ, rofyeP, and yafK) are neighboring genes to other genes that were found here to improve growth in ethanol, suggesting that they were enriched simply due to their genomic location rather than their encoded function. Thus, 9 of 11 genes in distinct clones were confirmed to confer ethanol tolerance.

The nine tolerant clones contain genes that have not previously been shown to improve growth in the presence of ethanol stress to our knowledge. These genes encode functions involved in a variety of cellular processes, such as: LPS biosynthesis (lpcA, arnB, and arnC), translation (tilS, yaeJ), fatty acid oxidation (fadE), serine biosynthesis (serA), regulation of transcription (zur), and transport (yicE) (Table 2-2).
<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>Gene product(s)</th>
<th>Size (bp)</th>
<th>Genomic position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 lpcA</td>
<td>D-sedoheptulose 7-phosphate isomerase</td>
<td>900</td>
<td>243,310 - 244,187</td>
</tr>
<tr>
<td>2 tilS</td>
<td>tRNAAla-lysidine synthetase</td>
<td>1,400</td>
<td>212,241 - 213,650</td>
</tr>
<tr>
<td>3 yhfT,</td>
<td>predicted inner membrane protein, predicted protein</td>
<td>1,900</td>
<td>3,504,054 - 3,505,915</td>
</tr>
<tr>
<td>yhfU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 fadE</td>
<td>acyl-CoA dehydrogenase</td>
<td>2,600</td>
<td>240,859 - 243,500</td>
</tr>
<tr>
<td>5 serA</td>
<td>α-ketoglutarate reductase / D-3-phosphoglycerate dehydrogenase</td>
<td>1,500</td>
<td>3,055,200 - 3,056,698</td>
</tr>
<tr>
<td>6 zur</td>
<td>Zur transcriptional repressor</td>
<td>900</td>
<td>4,257,501 - 4,258,373</td>
</tr>
<tr>
<td>7 arnC,</td>
<td>undecaprenyl phosphate-L-Ara4FN transferase, UDP-L-Ara4O C-4 transaminase</td>
<td>2,400</td>
<td>2,363,701 - 2,366,090</td>
</tr>
<tr>
<td>arnB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 yicE</td>
<td>YicE NCS2 transporter</td>
<td>1,700</td>
<td>3,826,723 - 3,828,374</td>
</tr>
<tr>
<td>9 yaeQ,</td>
<td>conserved protein, peptidyl-tRNA hydrolase (ribosome rescue factor)</td>
<td>1,100</td>
<td>214,150 - 215,255</td>
</tr>
<tr>
<td>yaeJ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 yafK</td>
<td>conserved protein</td>
<td>900</td>
<td>245,065 - 245,960</td>
</tr>
<tr>
<td>11 rof,</td>
<td>modulator of Rho-dependent transcription termination, conserved protein</td>
<td>600</td>
<td>245,065 - 245,960</td>
</tr>
<tr>
<td>yaeP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 yafJ</td>
<td>predicted amidotransferase</td>
<td>1,100</td>
<td>244,009 - 245,091</td>
</tr>
<tr>
<td>13 trmA</td>
<td>tRNA(m5U54)methyltransferase</td>
<td>1,200</td>
<td>4,160,193 - 4,161,397</td>
</tr>
<tr>
<td>14 yijD,</td>
<td>conserved inner membrane protein, FabR DNA-binding transcriptional repressor</td>
<td>1,200</td>
<td>4,159,003 - 4,160,208</td>
</tr>
<tr>
<td>fabR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2-2** Constructed clones containing highest-fitness genes identified from ethanol tolerance selections.

Although these genes have not been shown to directly improve ethanol tolerance previously, many of these processes have been linked to ethanol tolerance in *E. coli*. Ethanol is known to increase the permeability of the cell membrane, presumably by lodging within the hydrophobic core region of the lipid bilayer (Ingram, 1976; Ingram, 1990; Ingram and Buttke, 1984), and ethanol tolerance has been improved by various strategies to alter the fatty acid composition (Ingram, 1990; Ingram et al., 1980; Luo et al., 2009; Uchida, 1974). Therefore, we expected to find genes with functions related to the cell membrane composition that might combat the effects of ethanol or prevent its entry into the cell. Indeed, four of the nine confirmed genes encode functions related to membrane composition (fadE) or passage across the outer cell membrane (LPS biosynthesis via lpcA, arnB, arnC). Serine supplementation in LB medium has previously been shown to improve growth in ethanol (Gonzalez et al., 2003), and here we found that SerA,
the first committed step in serine biosynthesis, improves growth in the presence of ethanol. In the case of translation and transcription related genes, it is known that conditions that affect the function of the cell membrane under aerobic conditions lead to disruption of electron transport chain functions, with subsequent effects on transcription and translation (Kohanski et al., 2008; Spindler et al., 2011). While it is possible to rationalize how the functions of these confirmed tolerance genes fit into the existing understanding of ethanol toxicity, extensive follow up studies would allow for the elucidation of precise tolerance mechanisms and the extent to which such mechanisms overlap across our own studies and those of others.

For example, Goodarzi et al. performed ethanol selections in complex medium at mild and harsh ethanol concentrations using both overexpression libraries and transposon mutagenesis libraries (2010). They found enrichment in the overexpression libraries for the heat-shock stress response and the pathways involving propionate catabolism, glycine cleavage, glycine-betaine synthesis, and cell-wall biogenesis. We compared our data to their data using the same method for calculating gene fitness z-scores as they reported. Interestingly, the genes with highest fitness in each study differ substantially (Spearman rank coefficients: high ethanol= −0.03, moderate ethanol= 0.02) (Figure 2-4).
Comparison of gene fitness values reported here and previously by Goodarzi, et al (2010). Our MOPS minimal media selections were carried out at ethanol concentrations of 15 g/L (moderate) and 30 g/L (high), whereas Goodarzi et al. (2010) performed selections in LB rich media at 32 g/L (moderate) and 40 g/L ethanol (high).

We expect that the use of complex medium (LB) by Goodarzi et al compared to our use of minimal medium was the major contributor to such differences. Mutations which confer isobutanol tolerance in either a minimal media or a yeast extract supplemented media were found to differ significantly in previous studies (Minty et al., 2011; Smith and Liao, 2011). Specifically, the shift from rich to a minimal medium is well established to cause large metabolic shifts, particularly related to macromolecular synthesis and the availability of monomer pools (i.e. amino-acids). For example, in LB medium one would expect to select for genes involved in serine uptake, as opposed to the serine biosynthesis gene (serA) that we identified via minimal
medium selections. Genes responsible for ethanol degradation that are in the propionate catabolism pathway or have alcohol dehydrogenase activity were not enriched in our selections, nor did we have enrichment for peptidoglycan or glycine-betaine biosynthesis genes. It is plausible that such mechanisms exhibit the greatest benefit only in rich medium. Despite these differences, however, \textit{arnB} and \textit{arnC} were within the top 50 genes for the 4\% v/v selection by Goodarzi et al. (2010), and additionally, the highest fitness gene in both of their 4\% and 5.5\% v/v selections, the succinate semialdehyde reductase \textit{yihU} which can also reduce toxic accumulation of methylglyoxal (Saito et al., 2009), was moderately enriched in our moderate and high ethanol stress ($W_{yihU} \approx 1.1$ in both).

2.4.3 Transcriptional and Proteomic analysis of ethanol-tolerant clones

Using our SCALEs approach, we identified many genes for which overexpression appeared to increase growth in the presence of ethanol, nine of which were subsequently confirmed to enhance tolerance. In order to direct future mechanistic studies we next sought to better understand the extent to which overexpression of each of these genes resulted in the activation of similar pathways that contributed to the overall ethanol tolerance phenotype. That is, we speculated that while we employed different individual genetic strategies to confer tolerance, the resulting global expression phenotype may have been shared among some of these clones (Alper et al., 2006; Gill et al., 2002), and thus such clones would be more likely to confer tolerance through similar biochemical and/or physiological mechanisms. To explore this notion, we performed transcriptional profiling under ethanol stress for the 5 most improved ethanol-tolerant clones and the parent strain that contains the empty vector. Briefly, each of the five clones and the control strain were grown identically in MOPS minimal medium containing 25 g/L ethanol to mid-exponential phase (5 hrs), at which point samples were taken for RNA purification and
expression profiling via Affymetrix gene chips. In that there is conflicting data regarding the relationship between mRNA and protein levels, we also performed iTRAQ proteomic analyses on the two most tolerance clones as a further level of investigation of the extent of overlap in expression phenotypes.

Hierarchical clustering of the transcriptional profiling data indicated that the expression profiles separate into two main branches. The \textit{tilS} and \textit{lpcA} clones form one cluster, with global expression profiles most similar to the parent strain expression profile. The \textit{yhfUT}, \textit{fadE}, and \textit{serA} clones form a separate cluster (Figure 2-5a). These data suggest that the tolerance phenotype conferred by \textit{serA} requires the largest overall re-wiring of gene expression. This is particularly interesting when compared to the level of ethanol tolerance conferred, which is the lowest for \textit{serA} overexpression compared to the other four clones. In fact, our data show an inverse correlation between the level of tolerance conferred and the extent of gene expression changes for each clone (compare Figure 2-5A with Figure 2-3B). We would note here that while this is an interesting observation, and one worthy of additional investigation, it extends far beyond the intended scope of this particular study. Additionally, we performed clustering to compare the expression profiles of our ethanol-tolerant clones to expression profiles for ethanol-tolerant strains reported in previous studies by either adaptation (Horinouchi et al., 2010) or global transcription machinery engineering (Alper and Stephanopoulos, 2007). The adapted strains showed the most highly similar expression profiles, while gene overexpression-based tolerance phenotypes had slightly greater expression differences, and clones containing mutated housekeeping sigma factors (\textit{rpoD}) showed the greatest expression profile differences, including large differences between the ethanol-stressed and unstressed transcriptomes (Figure 2-6).
**Figure 2-5** Ethanol tolerance transcriptomes. (a) Transcriptomes of ethanol-tolerant clones. Hierarchical clustering of the ethanol-stressed transcriptomes based on the concentration of each gene’s mRNA concentration (Pearson’s correlation rows and columns and average linkage) shows strong similarity between biological replicates and two main clusters for the clones’ transcriptomes. Figure generated using GenePattern. (b-e) Venn diagrams generated using VennMaster for (b) overexpressed and (c) underexpressed genes (1.5-fold average) for each ethanol-tolerant clone compared to the parent strain. Gene expression changes were mapped to GO terms to evaluate significant (d) enrichment and (e) depletion of GO terms for each ethanol-tolerant clone.
Hierarchical clustering was performed to compare the altered transcriptomes of the ethanol-tolerant clones engineered in this study to those engineered in previous studies, specifically ethanol-tolerant clones containing mutant rpoD sigma factors engineered by 3 rounds of gTME (Alper et al., 2007 under GEO accession number GSE3665) and ethanol-tolerant clones selected by parallel adaptation to 5% (v/v) (~40 g/L) ethanol for 2,500 hours (Horinouchi et al., 2010 in supplementary table 1). Absent or zero-containing rows were eliminated for this analysis. Clustering was performed over a total of 2,145 genes.

Figure 2-6 Clustering of expression microarrays in ethanol tolerance engineering studies. Hierarchical clustering was performed to compare the altered transcriptomes of the ethanol-tolerant clones engineered in this study to those engineered in previous studies, specifically ethanol-tolerant clones containing mutant rpoD sigma factors engineered by 3 rounds of gTME (Alper et al., 2007 under GEO accession number GSE3665) and ethanol-tolerant clones selected by parallel adaptation to 5% (v/v) (~40 g/L) ethanol for 2,500 hours (Horinouchi et al., 2010 in supplementary table 1). Absent or zero-containing rows were eliminated for this analysis. Clustering was performed over a total of 2,145 genes.
In addition to hierarchical clustering, which is based on similarities among global gene expression changes, we were also interested in examining overlap in expression changes at the gene function level. To do so, we compiled lists of the genes that had at least a 1.5-fold increase (Figure 2-5B) or decrease in expression (Figure 2-5C) on average for each clone and then mapped these gene lists of expression changes to gene ontology categories (called ‘GO terms’) using GOMiner (Zeeberg et al., 2003), which allowed us to identify for statistically significant representations of any GO terms within the group of genes with increased or decreased expression changes. We noticed that the similarities between expression patterns of these clones seen through clustering are also reflected in the overlap of the lists of overexpressed (Figure 2-5B) and to a lesser extent for the underexpressed genes (Figure 2-5C). The lpcA clone shares the highest overlap with tilS for expression changes. For the yhfUT, fadE, and serA clones, all three largely intersect for the overexpressed genes with yhfUT being almost entirely a subset (~80%) of the fadE and serA overexpressed genes, but the overlap does not closely match the clustering for the underexpression changes.

Not surprisingly, there is similar overlap once we map the overexpressed and underexpressed genes for each clone to GO terms respectively (Figure 2-5D, Figure 2-5E). Interestingly, the vast majority (20 out of 26) of the GO terms overrepresented among the overexpressed genes in all 5 ethanol-tolerant clones are related to transcription regulation and regulation of biological processes (Table 2-3, there are a total of 3,561GO terms). This observation supports the finding that engineering transcription network components, either directly as performed by Alper and Stephanopoulos (2007) or indirectly as suggested here, could be an effective approach for engineering ethanol tolerance.
We also performed this analysis for the 2-fold expression changes for each clone, but this cutoff was too stringent as the enriched GO terms (p < 0.05) were dominated by highly specific GO terms with only a few gene members (median number of total genes per GO term dropped from 23 to 3 gene members for 1.5-fold versus 2-fold expression changes). Some other interesting upregulated GO terms that are unique to certain clones include: upregulation in the lpcA clone of

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO6350:</td>
<td>transcription</td>
</tr>
<tr>
<td>GO6351:</td>
<td>transcription, DNA-dependent</td>
</tr>
<tr>
<td>GO6355:</td>
<td>regulation of transcription, DNA-dependent</td>
</tr>
<tr>
<td>GO9889:</td>
<td>regulation of biosynthetic process</td>
</tr>
<tr>
<td>GO19219:</td>
<td>regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process</td>
</tr>
<tr>
<td>GO10468:</td>
<td>regulation of gene expression</td>
</tr>
<tr>
<td>GO10556:</td>
<td>regulation of macromolecule biosynthetic process</td>
</tr>
<tr>
<td>GO19222:</td>
<td>regulation of metabolic process</td>
</tr>
<tr>
<td>GO31323:</td>
<td>regulation of cellular metabolic process</td>
</tr>
<tr>
<td>GO31326:</td>
<td>regulation of cellular biosynthetic process</td>
</tr>
<tr>
<td>GO32774:</td>
<td>RNA biosynthetic process</td>
</tr>
<tr>
<td>GO45449:</td>
<td>regulation of transcription</td>
</tr>
<tr>
<td>GO50789:</td>
<td>regulation of biological process</td>
</tr>
<tr>
<td>GO50794:</td>
<td>regulation of cellular process</td>
</tr>
<tr>
<td>GO51171:</td>
<td>regulation of nitrogen compound metabolic process</td>
</tr>
<tr>
<td>GO51252:</td>
<td>regulation of RNA metabolic process</td>
</tr>
<tr>
<td>GO60255:</td>
<td>regulation of macromolecule metabolic process</td>
</tr>
<tr>
<td>GO65007:</td>
<td>biological regulation</td>
</tr>
<tr>
<td>GO80090:</td>
<td>regulation of primary metabolic process</td>
</tr>
<tr>
<td>GO2000112:</td>
<td>regulation of cellular macromolecule biosynthetic process</td>
</tr>
<tr>
<td>GO5622:</td>
<td>intracellular</td>
</tr>
<tr>
<td>GO44424:</td>
<td>intracellular part</td>
</tr>
<tr>
<td>GO9318:</td>
<td>exodeoxyribonuclease VII complex</td>
</tr>
<tr>
<td>GO8855:</td>
<td>exodeoxyribonuclease VII activity</td>
</tr>
<tr>
<td>GO6308:</td>
<td>DNA catabolic process</td>
</tr>
<tr>
<td>GO5737:</td>
<td>cytoplasm</td>
</tr>
</tbody>
</table>

Table 2-3 Enriched Up GO terms for all clones determined from transcriptional profiling data. List of all GO terms enriched (p< 0.05) for the 1.5-fold overexpressed genes in all clones
the glycine cleavage complex (GO:0005960), which has previously been shown to boost ethanol tolerance (Gonzalez et al., 2003; Goodarzi et al., 2010); upregulation in the tilS clone of iron ion homeostasis and binding (GO:0006879, GO:0008199, GO:0055072), where iron is known to play a role in hydroxyl radical formation (Liochev, 1996), we have measured increased hydroxyl radical production under similar ethanol stress (unpublished data), and iron ion homeostasis has previously been found to be upregulated in ethanol-tolerant clones (Horinouchi et al., 2010) and perturbed during isobutanol stress (Brynildsen and Liao, 2009); and upregulation in the serA clone of vitamin B6 biosynthesis (GO:0042819, GO:0042816), which may play a protective role against oxidative stress here as has been shown in other species (Bilski et al., 2000; Graham et al., 2004; Mooney et al., 2009). Additionally, the terms for response to stress and response to heat (GO:0006950, GO:0009408) are downregulated in both the lpcA and serA clones, which is in contrast to previous studies that found a correlation between ethanol tolerance and stress response overexpression (Gonzalez et al., 2003; Goodarzi et al., 2010).

For similar reasons as for the transcriptional profiling studies, and to provide additional insights that might exist at the protein but not mRNA levels, we performed a global proteomic analysis for the two most ethanol-tolerant clones and the parent empty vector strain (the three which formed the most similar cluster in Figure 2-5A) using isobaric tag labeling. For the lpcA clone there were 17 upregulated and 11 downregulated proteins (Table 2-4), and for the tilS clone, there were 22 upregulated and 17 downregulated proteins (Table 2-5) (P <0.05 for all) as compared to the parent strain (a total of 616 total proteins were quantified). LpcA and TilS were in the most highly upregulated proteins for each clone (approximately 2-fold and 4-fold increase), respectively.
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>lpcA/control concentration</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ymfP</em></td>
<td>0.36</td>
<td>Putative protein YmfP (Uncharacterized protein YmfP in lambdoid prophage e14 region)</td>
</tr>
<tr>
<td><em>msrA</em></td>
<td>0.47</td>
<td>Peptide methionine sulfoxide reductase MsrA (Protein-methionine-(\text{S})-oxide reductase) (EC 1.8.4.11)</td>
</tr>
<tr>
<td><em>pspB</em></td>
<td>0.56</td>
<td>Phage shock protein B</td>
</tr>
<tr>
<td><em>yaeH</em></td>
<td>0.62</td>
<td>UPF0325 protein yaeH</td>
</tr>
<tr>
<td><em>rpmD</em></td>
<td>0.62</td>
<td>50S ribosomal protein L30</td>
</tr>
<tr>
<td><em>metN</em></td>
<td>0.64</td>
<td>Methionine import ATP-binding protein MetN (EC 3.6.3.-)</td>
</tr>
<tr>
<td><em>menB</em></td>
<td>0.66</td>
<td>1,4-Dihydroxy-2-naphthoyl-CoA synthase (DHNA-CoA synthase) (EC 4.1.3.36)</td>
</tr>
<tr>
<td><em>degP</em></td>
<td>0.68</td>
<td>Protease do (EC 3.4.21.-)</td>
</tr>
<tr>
<td><em>gpt</em></td>
<td>0.74</td>
<td>Xanthine phosphoribosyltransferase (EC 2.4.2.22)</td>
</tr>
<tr>
<td><em>eno</em></td>
<td>0.77</td>
<td>Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (2-phosphoglycerate dehydratase)</td>
</tr>
<tr>
<td><em>tufA</em></td>
<td>0.89</td>
<td>Elongation factor Tu 1 (EF-Tu 1) (P-43)</td>
</tr>
<tr>
<td><em>rplL</em></td>
<td>1.19</td>
<td>50S ribosomal protein L9</td>
</tr>
<tr>
<td><em>aspC</em></td>
<td>1.23</td>
<td>Aspartate aminotransferase (AspAT) (EC 2.6.1.1) (Transaminase A)</td>
</tr>
<tr>
<td><em>hns</em></td>
<td>1.28</td>
<td>DNA-binding protein H-NS (Histone-like protein HLP-II) (Protein B1) (Protein H1)</td>
</tr>
<tr>
<td><em>ychF</em></td>
<td>1.29</td>
<td>GTP-dependent nucleic acid-binding protein engD</td>
</tr>
<tr>
<td><em>rpmJ</em></td>
<td>1.35</td>
<td>50S ribosomal protein L36 (Ribosomal protein B)</td>
</tr>
<tr>
<td><em>rpsH</em></td>
<td>1.37</td>
<td>30S ribosomal protein S8</td>
</tr>
<tr>
<td><em>rplB</em></td>
<td>1.44</td>
<td>50S ribosomal protein L2</td>
</tr>
<tr>
<td><em>sodB</em></td>
<td>1.45</td>
<td>Superoxide dismutase [Fe] (EC 1.15.1.1)</td>
</tr>
<tr>
<td><em>ilvB</em></td>
<td>1.53</td>
<td>Acetolactate synthase isozyme 1 large subunit (AHAS-I) (EC 2.2.1.6)</td>
</tr>
<tr>
<td><em>rplT</em></td>
<td>1.59</td>
<td>50S ribosomal protein L20</td>
</tr>
<tr>
<td><em>rplV</em></td>
<td>1.64</td>
<td>50S ribosomal protein L22</td>
</tr>
<tr>
<td><em>aceA</em></td>
<td>1.72</td>
<td>Isocitrate lyase (ICL) (Isocitrato) (Isocitratase) (EC 4.1.3.1)</td>
</tr>
<tr>
<td><em>pdxK</em></td>
<td>2.07</td>
<td>Pyridoxine kinase (EC 2.7.1.35) (PN/PL/PM kinase) (Vitamin B6 kinase)</td>
</tr>
<tr>
<td><em>lpcA</em></td>
<td>2.35</td>
<td>Phosphoheptose isomerase (EC 5.3.1.28) (Sedoheptulose 7-phosphate isomerase)</td>
</tr>
<tr>
<td><em>fabD</em></td>
<td>2.64</td>
<td>Malonyl CoA-acyl carrier protein transacylase (MCT) (EC 2.3.1.39)</td>
</tr>
<tr>
<td><em>mgtA</em></td>
<td>5.78</td>
<td>Magnesium-transporting ATPase, P-type 1 (EC 3.6.3.2) (Mg(2+) transport ATPase, P-type 1)</td>
</tr>
<tr>
<td><em>yadD</em></td>
<td>6.48</td>
<td>Uncharacterized protein yadD</td>
</tr>
</tbody>
</table>

Table 2-4 Significantly up- or down-regulated proteins in lpcA clone (p< 0.05) determined by iTRAQ.
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>tilS/control concentration</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ympP</td>
<td>0.40</td>
<td>Putative protein YmpP (Uncharacterized protein YmpP in lambdoid prophage e14 region)</td>
</tr>
<tr>
<td>ibpA</td>
<td>0.44</td>
<td>Small heat shock protein ibpA (16 kDa heat shock protein A)</td>
</tr>
<tr>
<td>fumA</td>
<td>0.65</td>
<td>Fumarate hydratase class I, aerobic (Fumarase) (EC 4.2.1.2)</td>
</tr>
<tr>
<td>glgB</td>
<td>0.66</td>
<td>1,4-alpha-glucan-branching enzyme (EC 2.4.1.18) (Glycogen-branching enzyme) (BE)</td>
</tr>
<tr>
<td>yicl</td>
<td>0.68</td>
<td>Alpha-xilosidase (EC 3.2.1.-)</td>
</tr>
<tr>
<td>gatA</td>
<td>0.69</td>
<td>Galactitol-specific phosphotransferase enzyme IIA component (EC 2.7.1.-)</td>
</tr>
<tr>
<td>hinT</td>
<td>0.70</td>
<td>HIT-like protein hinT</td>
</tr>
<tr>
<td>bamA</td>
<td>0.72</td>
<td>Outer membrane protein assembly factor yaeT (Omp85)</td>
</tr>
<tr>
<td>aldA</td>
<td>0.73</td>
<td>Lactaldehyde dehydrogenase (EC 1.2.1.22) (Glycolaldehyde dehydrogenase) (EC 1.2.1.21)</td>
</tr>
<tr>
<td>vigA</td>
<td>0.74</td>
<td>Uncharacterized protein vigA</td>
</tr>
<tr>
<td>sdhA</td>
<td>0.74</td>
<td>Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1)</td>
</tr>
<tr>
<td>ygbD</td>
<td>0.74</td>
<td>Alcohol dehydrogenase YgbD (EC 1.1.1.-)</td>
</tr>
<tr>
<td>cysK</td>
<td>0.76</td>
<td>Cysteine synthase A (CSase A) (EC 2.5.1.47) (Sulfate starvation-induced protein 5) (SSI5)</td>
</tr>
<tr>
<td>rpsS</td>
<td>0.77</td>
<td>30S ribosomal protein S19</td>
</tr>
<tr>
<td>bfr</td>
<td>0.77</td>
<td>Bacterioferritin (BFR) (Cytochrome b-1) (Cytochrome b-557)</td>
</tr>
<tr>
<td>yfeX</td>
<td>0.79</td>
<td>Uncharacterized protein yfeX (EC 1.-...-.)</td>
</tr>
<tr>
<td>dnaK</td>
<td>0.81</td>
<td>Chaperone protein DnaK (HSP70) (Heat shock 70 kDa protein) (Heat shock protein 70)</td>
</tr>
<tr>
<td>aceE</td>
<td>1.21</td>
<td>Pyruvate dehydrogenase E1 component (EC 1.2.4.1)</td>
</tr>
<tr>
<td>talB</td>
<td>1.23</td>
<td>Transaldolase B (EC 2.2.1.2)</td>
</tr>
<tr>
<td>alaS</td>
<td>1.23</td>
<td>Alanyl-tRNA synthetase (EC 6.1.1.7) (Alanine--tRNA ligase) ( AlaRS)</td>
</tr>
<tr>
<td>hns</td>
<td>1.25</td>
<td>DNA-binding protein H-NS (Histone-like protein HLP-II) (Protein B1) (Protein H1)</td>
</tr>
<tr>
<td>ppsA</td>
<td>1.28</td>
<td>Phosphoenolpyruvate synthase (PEP synthase) (EC 2.7.9.2) (Pyruvate, water dikinase)</td>
</tr>
<tr>
<td>thrC</td>
<td>1.28</td>
<td>Threonine synthase (TS) (EC 4.2.3.1)</td>
</tr>
<tr>
<td>fabI</td>
<td>1.29</td>
<td>Enoyl-[acyl-carrier-protein] reductase [NADH] (EC 1.3.1.9)</td>
</tr>
<tr>
<td>purC</td>
<td>1.30</td>
<td>Phosphoribosylaminomimidazole-succinocarboxamide synthase (EC 6.3.2.6)</td>
</tr>
<tr>
<td>sodB</td>
<td>1.30</td>
<td>Superoxide dismutase [Fe] (EC 1.15.1.1)</td>
</tr>
<tr>
<td>crr</td>
<td>1.31</td>
<td>Glucose-specific phosphotransferase enzyme IIA component (EC 2.7.1.-)</td>
</tr>
<tr>
<td>glnA</td>
<td>1.33</td>
<td>Glutamine synthetase (EC 6.3.1.2) (Glutamate--ammonia ligase)</td>
</tr>
<tr>
<td>rpmJ</td>
<td>1.43</td>
<td>50S ribosomal protein L36 (Ribosomal protein B)</td>
</tr>
<tr>
<td>rplE</td>
<td>1.43</td>
<td>50S ribosomal protein L5</td>
</tr>
<tr>
<td>mdoG</td>
<td>1.50</td>
<td>Glucans biosynthesis protein G</td>
</tr>
<tr>
<td>glnS</td>
<td>1.51</td>
<td>Glutaminyl-tRNA synthetase (EC 6.1.1.18) (Glutamine--tRNA ligase) (GlnRS)</td>
</tr>
<tr>
<td>rfaD</td>
<td>1.53</td>
<td>ADP-L-glycero-D-manno-heptose-6-epimerase (EC 5.1.3.20)</td>
</tr>
<tr>
<td>pepA</td>
<td>1.64</td>
<td>Cytosol aminopeptidase (EC 3.4.11.1) (Aminopeptidase A/I) (Leucine aminopeptidase)</td>
</tr>
<tr>
<td>bcp</td>
<td>1.81</td>
<td>Putative peroxiredoxin bcp (EC 1.11.1.15) (Thioredoxin reductase)</td>
</tr>
<tr>
<td>cydD</td>
<td>1.83</td>
<td>ATP-binding/permease protein CydD</td>
</tr>
<tr>
<td>pdxK</td>
<td>1.85</td>
<td>Pyridoxine kinase (EC 2.7.1.35) (Vitamin B6 kinase)</td>
</tr>
<tr>
<td>glgA</td>
<td>2.06</td>
<td>Glycogen synthase (EC 2.4.1.21) (Starch [bacterial glycogen] synthase)</td>
</tr>
<tr>
<td>tilS</td>
<td>3.89</td>
<td>tRNA(Ile)-lysidine synthase (EC 6.3.4.-)</td>
</tr>
</tbody>
</table>

Table 2-5 Significantly up- or down-regulated proteins in tilS clone (p< 0.05) determined by iTRAQ.
This investigation provided several interesting outcomes. First, there is relatively little in common between the gene subsets that are up- or down-regulated at the protein level or mRNA level (Figure 2-7A), and there is a weakly positive correlation between the change in concentration of the mRNA and protein for these two clones as compared to the parent with Spearman rank correlation coefficients of 0.11 and 0.35 for the \( lpcA \) and \( tilS \) clones respectively (Figure 2-8).

These data support numerous prior studies suggesting that changes at the mRNA level are not always reflected at the protein level (Foss et al., 2011; Gygi et al., 1999; Lu et al., 2007). Second, there is significant overlap for the enriched GO terms in the gene lists of the upregulated proteins for the \( lpcA \) and \( tilS \) clones (Figure 2-7B). Both clones have enrichment for GO terms related to superoxide radical mitigation, vitamin B6 biosynthesis, translation, and fatty acid
biosynthesis (Table 2-6). As expected, these GO terms differ from those found in the mRNA studies (transcription and regulation not enriched as noted above).

Figure 2-8 Comparison of expression changes at RNA and protein level for lpcA and tilS clones. The average fold change from transcriptional profiling and proteomics for upregulated protein lists of lpcA clone (blue diamonds) and tilS clone (red squares).

In summary, our analysis here suggested that indeed there appears to be overlap in the global expression changes underlying the ethanol tolerance phenotypes of each evaluated clone with 39 % of overexpressed genes (215/545 genes, 19 % in 2 clones, 9 % in 3 clones, 5 % in 4 clones, and 6 % in 5 clones) and 28 % of underexpressed genes (43/154 genes, 19 % in 2 clones, 5 % in 3 clones, 3 % in 4 clones, 1 % in 5 clones) having similarly altered expression in multiple ethanol-tolerant clones. Nevertheless, the extent of overlap differed for each clone at both the level of individual gene transcripts or proteins as well as when summarized into GO terms.
<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO303</td>
<td>response to superoxide</td>
</tr>
<tr>
<td>GO305</td>
<td>response to oxygen radical</td>
</tr>
<tr>
<td>GO4784</td>
<td>superoxide dismutase activity</td>
</tr>
<tr>
<td>GO16721</td>
<td>oxidoreductase activity, acting on superoxide radicals as acceptor</td>
</tr>
<tr>
<td>GO19430</td>
<td>removal of superoxide radicals</td>
</tr>
<tr>
<td>GO6801</td>
<td>superoxide metabolic process</td>
</tr>
<tr>
<td>GO71450</td>
<td>cellular response to oxygen radical</td>
</tr>
<tr>
<td>GO71451</td>
<td>cellular response to superoxide</td>
</tr>
<tr>
<td>GO3681</td>
<td>bent DNA binding</td>
</tr>
<tr>
<td>GO5488</td>
<td>binding</td>
</tr>
<tr>
<td>GO3735</td>
<td>structural constituent of ribosome</td>
</tr>
<tr>
<td>GO5840</td>
<td>ribosome</td>
</tr>
<tr>
<td>GO6412</td>
<td>translation</td>
</tr>
<tr>
<td>GO19538</td>
<td>protein metabolic process</td>
</tr>
<tr>
<td>GO30529</td>
<td>ribonucleoprotein complex</td>
</tr>
<tr>
<td>GO34645</td>
<td>cellular macromolecule biosynthetic process</td>
</tr>
<tr>
<td>GO44237</td>
<td>cellular metabolic process</td>
</tr>
<tr>
<td>GO44238</td>
<td>primary metabolic process</td>
</tr>
<tr>
<td>GO44249</td>
<td>cellular biosynthetic process</td>
</tr>
<tr>
<td>GO44267</td>
<td>cellular protein metabolic process</td>
</tr>
<tr>
<td>GO42816</td>
<td>vitamin B6 metabolic process</td>
</tr>
<tr>
<td>GO42819</td>
<td>vitamin B6 biosynthetic process</td>
</tr>
<tr>
<td>GO4312</td>
<td>fatty acid synthase activity</td>
</tr>
<tr>
<td>GO4340</td>
<td>glucokinase activity</td>
</tr>
<tr>
<td>GO4396</td>
<td>hexokinase activity</td>
</tr>
<tr>
<td>GO4743</td>
<td>pyruvate kinase activity</td>
</tr>
<tr>
<td>GO44424</td>
<td>intracellular part</td>
</tr>
<tr>
<td>GO44444</td>
<td>cytoplasmic part</td>
</tr>
<tr>
<td>GO5622</td>
<td>intracellular</td>
</tr>
<tr>
<td>GO5737</td>
<td>cytoplasm</td>
</tr>
<tr>
<td>GO5829</td>
<td>cytosol</td>
</tr>
<tr>
<td>GO8152</td>
<td>metabolic process</td>
</tr>
<tr>
<td>GO8478</td>
<td>pyridoxal kinase activity</td>
</tr>
<tr>
<td>GO8614</td>
<td>pyridoxine metabolic process</td>
</tr>
<tr>
<td>GO8615</td>
<td>pyridoxine biosynthetic process</td>
</tr>
<tr>
<td>GO9058</td>
<td>biosynthetic process</td>
</tr>
<tr>
<td>GO9059</td>
<td>macromolecule biosynthetic process</td>
</tr>
</tbody>
</table>

**Table 2-6** Upregulated GO terms for lpcA and tilS under ethanol stress from iTRAQ proteomic analysis. List of all GO terms enriched (p< 0.05) for the significantly upregulated proteins.
2.5 Conclusions

In this work, we have used a genome-wide approach to identify many individual genes from a broad array of biological processes that individually improve the growth of *E. coli* under ethanol stress. To the best of our knowledge, these genes have not previously been identified for ethanol tolerance, and they confer up to an approximately six-fold improvement in growth with high ethanol stress. Through transcriptional profiling and proteomic analysis, we found that some of the ethanol-tolerant clones appear to manipulate the global expression landscape in a similar manner. The level of similarity differed both at the level of each clone, as well as when comparing changes in mRNA versus protein expression. Interestingly, when expression data was summarized at the functional level (i.e. GO terms), we found a substantial enrichment for regulatory functions in the set of genes with increased expression across all five clones analyzed, while protein expression changes did not exhibit the same level of functional coordination. These data suggest that further engineering at the level of regulation of expression, but targeted towards regulons involved in tolerance phenotypes such as oxygen radical mitigation, may be a worthwhile future endeavor. Overall, while the methods that we have applied here are applicable to many different traits, the specific genetic targets that we have identified could be useful for future efforts to engineer tolerance to other short-chain alcohols or further improve ethanol tolerant strains through combinatorial optimization.
Acknowledgments

We thank Nich Sandoval and Eileen Spindler for helpful discussions on the SCALEs analysis and Helen Marshall at the Microarray Facility for running the arrays. LBA Woodruff was supported by a NSF Graduate Research Supplement (CBET0449183). We thank the EPSRC (EP/E036252/1) and the Bioprocess Research Industry Club (BRIC) (BB/F004842/1) for funding the proteomics work.

2.6 References


Chapter 3 Towards a metabolic engineering strain “commons”:

An *Escherichia coli* platform strain for ethanol production

Accepted pending revision at *Biotechnology and Bioengineering* (November 2012).

3.1 Abstract

In the genome-engineering era it is increasingly important that researchers have access to a common set of platform strains that can serve as debugged production chassis and the basis for applying new metabolic engineering strategies for modeling and characterizing flux, engineering complex traits, and optimizing overall performance. Here we describe such a platform strain of *E. coli* engineered for ethanol production. Starting with a fully characterized host strain (BW25113), we site-specifically integrated the genes required for homoethanol production under the control of a strong inducible promoter into the genome and deleted the genes encoding four enzymes from competing pathways. This strain is capable of producing >30 g/L of ethanol in minimal media with < 2 g/L produced of any fermentative byproduct. Using this platform strain, we tested previously identified ethanol tolerance genes and found that while tolerance was improved under certain conditions, any effect on ethanol production or tolerance was lost when grown under production conditions. Thus, our findings reinforce the need for a metabolic engineering “commons” that could provide a set of platform strains for use in more sophisticated genome-engineering strategies. Towards this end, we have made this production strain available to the scientific community through the American Type Culture Collection (ATCC).
3.2 Introduction

Recent advancements in genome-engineering approaches and capabilities have the potential to be coupled with the tremendous achievements in metabolic engineering and ultimately provide a better understanding of strategies for optimizing biological production of a desired chemical. Genome engineering tools, such as genome-scale models (Edwards and Palsson, 2000; Feist et al., 2007; Reed et al., 2003; Schilling et al., 1999), global transcription machinery engineering (gTME) (Alper et al., 2006; Alper and Stephanopoulos, 2007), multiplex automated genome engineering (MAGE) (Wang et al., 2009), trackable multiplex recombineering (TRMR) (Sandoval et al., 2012; Warner et al., 2010b), and other “omics” approaches, would benefit from a set of platform strains that could be applied as production chassis to investigate approaches to engineer desirable traits and would allow for direct comparisons to be made between studies. Although strains are routinely exchanged among academic groups (but not for-profit entities), this process is often encumbered by a need to negotiate individual agreements, a lack of staff to manage the exchange, and a lack of resources to maintain and verify the strain(s).

Many of the challenges encountered in synthetic biology when building systems from modules or parts, such as debugging and non-orthogonality (Arkin and Fletcher, 2006; Purnick and Weiss, 2009), are analogous to the metabolic engineering challenges involved with engineering a production strain. The steps of designing the genetic modifications, debugging to provide sufficient productivity and yield, and optimizing to achieve inhibitory product titer in a fully-genotyped production strain remains time-intensive (about 1 person-year for the LW06 strain here and for the succinate strains reported in Singh et al. (2009; 2011)) yet mostly requires the use of standard laboratory cloning strategies that in and of themselves do not routinely lead to
new insights. To date, the metabolic engineering community does not have a “commons” that might house an open-source collection of standardized production platform strains analogous to synthetic biology’s BioBricks Foundation and Registry of Standard Biological Parts, which facilitate the sharing of standardized DNA parts (Shetty et al., 2008). Here we use the term open source not solely to mean that the production strain is openly available (i.e., free-to-use) but also to indicate that the genetic code was written to enable further genome engineering and reconstruction (i.e., characterized and fully sequenced genetic modifications and genome). This proposed metabolic engineering commons repository of production platform strains could facilitate collaborative innovations, accelerate advancements in metabolic and genome engineering of production systems, and be used in conjunction with standard biological parts (i.e., BioBricks). As promoted by Keasling (2008), “Open-source biological components and eventually whole cells will decrease the cost of engineering metabolic pathways, make metabolic engineering more predictable with less guesswork, and encourage the development of novel biological solutions to some of our most challenging problems” (p. 65). Of course, this concept also builds upon long-standing culture collections, such as the American Type Culture Collection (ATCC) and Yale’s *E. coli* Genetic Stock Center (CGSC).

The proposed commons for production strains raises many issues (e.g., property rights of the depositor and user, standards for deposited strains) that should ultimately be resolved by the community. To initiate this dialogue, however, we offer our recommendations as a starting point. A large motivation for a single commons collection is to provide access to engineered strains without burdening individual labs to provide the material and negotiate the material transfer agreement (MTA) for each request. To this end, we recommend a third-party to maintain the commons collection and distribute the strains. For now, we have selected the
ATCC for this purpose; however, the standard ATCC agreements do not allow for commercial use of the provided material. In its current implementation, the commons provides an avenue for sharing production strains within the scientific community for non-commercial use. A universal commons agreement governing the deposit and transfer of strains in the commons would further simplify this process and could specifically address the commons policy regarding commercial use of the strains. For this purpose, we promote the use of the BioBrick Public Agreement (https://biobricks.org/bpa/), whereby materials are available to users free-to-use for non-commercial or commercial use while contributors have agreed not to assert property rights held against users. We have contacted the BioBricks Foundation towards this end. We expect that this will broaden innovation related to strain engineering and lead to advancements within the metabolic engineering community. Regarding the standards for deposited strains, we recommend that the strain have: (1) characterized and fully sequenced genetic modifications (i.e., the ability to fully reconstruct the entire genome), (2) inducible control of heterologous pathways, and (3) chromosomal integration of the production pathway. We also recommend that all strains deposited to the commons be characterized by providing data for production in minimal media with no supplementation (e.g., rate, titer, yield, and cell growth data). We used these recommendations to design our first entry into the commons as presented here.

As a step towards this longer term goal of a metabolic engineering commons, here we present the construction of an ethanol production platform strain, LW06, which we have made openly available through the ATCC (ATCC# BAA-2466). Ethanol production in *E. coli* is a potential route to producing renewable biofuels, and decades of research (primarily by Lonnie Ingram and colleagues) have been conducted to improve titers and substrate utilization, among other desirable traits in this system (Dien et al., 2000; Jarboe et al., 2007; Ohta et al., 1991). *E.
coli heterologously expressing the homoethanol pathway from Zymomonas mobilis has been shown to be an efficient approach to produce high yields of ethanol in an E. coli background (Alterthum and Ingram, 1989; Ohta et al., 1991), and the use E. coli as the host allows for easy genetic manipulation and a wide-range of genomic approaches to be used. The landmark patent for heterologous expression of the Z. mobilis ethanol production pathway (pdc and adhB) in E. coli was issued 21 years ago (Ingram et al., 1991) and therefore expired 4 years ago. Several studies have utilized ethanol production strains with plasmid-based expression of the homoethanol pathway (Alterthum and Ingram, 1989; Dien et al., 2000); however, chromosomal integration of the heterologous ethanol production cassette, such as those produced by Lonnie Ingram (Ohta et al., 1991; Yomano et al., 1998; Yomano et al., 2008), provides greater genetic stability over time and more versatility for using plasmid-based genomic approaches to identify trait-conferring genetic elements.

The LW06 ethanol production platform that we have constructed here differs from previously constructed strains in a few key aspects: (1) it contains inducible ethanol production, (2) the ethanol production cassette is inserted in an intergenic region with minimal polar effects, (3) it has not undergone any adaptation and/or screening, (4) it was not engineered using transposable elements that have been found to allow for genetic rearrangements (Turner et al., 2012), and (5) competing fermentative pathways have been knocked out, but no other genetic modifications to enhance production or for other traits were introduced. With these attributes, this ethanol production platform fulfills nearly all of the aspects of an ideal microbial factory chassis previously discussed (Keasling, 2008).
3.3 Results and Discussion

To construct our platform ethanol production strain, we started by cloning the homoethanol production pathway (\(pdc_{Zm}, adhB_{Zm}\) under \(P_{lac}\) promoter) from pLOI297 (ATCC 68239), adding an ampicillin-resistance cassette, and inserting this ethanol production cassette at the \(attTn7\) intergenic site (replacing base pair 3,909,796), which was previously determined to have minimal polar effects (Datsenko and Wanner, 2000; Warner et al., 2010b). We tested the activity of the ethanol pathway in the resultant LW01 strain on aldehyde indicator plates, which have previously been found to correlate to ethanol production (Conway et al., 1987), and found minimal activity as shown in Figure 3-1. We hypothesized that the decreased activity compared to the pLOI297 plasmid was due to the decreased copy number of the ethanol production pathway when integrated as compared to the 15 - 20 copies/cell of the pLOI297 plasmid. Therefore, we swapped the \(P_{lac}\) promoter for the more highly expressed and IPTG-inducible \(P_{LlacO-1}\), based on the \(P_L\) promoter of phage \(\lambda\) (Lutz and Bujard, 1997), to construct the LW02 strain. Upon testing LW02 with IPTG induction on aldehyde indicator plates, we found that LW02 has largely improved activity of the ethanol production pathway as shown in Figure 3-1. The amount of color change on the indicator plates for LW02 was equivalent to that achieved with pLOI297 and was found to be equivalent without IPTG induction, which suggests that despite a 600-fold increase in expression with IPTG induction (Lutz and Bujard, 1997), the leakiness of the promoter is enough to saturate the indicator plates (data not shown).
Ethanol production chromosomal integration and activity. The ethanol production cassette comprised of pdc and adhB from Z. mobilis under the control the P underscores{lac} promoter and an ampicillin-resistance marker were integrated into the chromosome at the attTn7, replacing base-pair 3,909,796, by λ Red-mediated homologous recombination using pSIM5 and resulting in strain LW01. The P underscores{lac} promoter was swapped for the P underscores{LlacO-1} to achieve higher activity of the ethanol production pathway. Activity of the ethanol production pathway was tested using aldehyde indicator plates (Conway et al., 1987) with 1 mM IPTG induction.

To quantitatively test the ethanol production of the LW02 strain, we performed batch fermentations shown in Figure 3-2A with the AM1 minimal production media (without betaine addition) previously optimized for ethanol production (Martinez et al., 2007). LW02 produced 16 g/L ethanol with 50 % of the theoretical yield due to the substantial accumulation of additional fermentative byproducts (Figure 3-2B).
To prevent this undesired accumulation of byproducts, we knocked out enzymes in the competing fermentative pathways for acetate (ackA), succinate (frdABCD), lactate (ldhA), and native ethanol production (adhE), which is not redox balanced. The pathway engineering strategy we employed here targets the same competing enzymes as used by Yomano et al. (2008). As shown in Figure 3-2C and Figure 3-2D, the final resulting strain, LW06, when grown...
in minimal media produced 28 g/L ethanol and less than 2 g/L for any of the fermentative byproducts, which is 80% of the theoretical yield based on the measured glucose consumption. Previous studies in AM1 media, however with betaine supplementation, and 90 g/L glucose or xylose have reported ethanol produced at 90 – 95% of the theoretical yield for engineered E. coli strains (Martinez et al. 2007; Yomano et al. 2009).

To demonstrate an application of the ethanol production platform as envisioned above, we tested whether individual genes for ethanol tolerance, which we previously identified and verified in a wild-type, non-production system, would improve ethanol tolerance and production in this strain. The most beneficial genes for E. coli tolerance to exogenous ethanol that we identified through our previously described genome-wide search (Woodruff et al, revisions submitted) encode for the enzyme that carries out the initial committed step in the biosynthesis of the core region of lipopolysaccharides (lpcA) and the tRNA\text{Ile}-lysidine synthetase that is responsible for the proper discrimination and tRNA charging of isoleucine versus the structurally similar methionine (tilS) (Ikeuchi et al., 2005). Interestingly, we found that production strains containing plasmid-based lpcA had no substantial effect on ethanol production while plasmid-based tilS reduced the final ethanol titer by approximately 30%, as shown in Figure 3-3A.
LpcA is involved in the biosynthesis of lipopolysaccharides that decorate the outer membrane and probably acts to alter the membrane hydrophobicity. As discussed by Dunlop in a recent review (2011), tolerance to exogenous inhibitory products (e.g., alcohol biofuels) is widely studied yet rarely in a system with intracellular production which might require different tolerance mechanisms (e.g., preventing entry of exogenous alcohol into the cell may similarly prevent endogenous product from diffusing out of the cell), and the common assumption is that tolerance will improve production, yet only some studies support this assumption (Alper et al.,...
2006; Dunlop et al., 2011; Yomano et al., 1998), while others dispute this assumption (Atsumi et al., 2010; Baer et al., 1987; Zhao et al., 2003). Therefore, we hypothesize that \textit{lpcA} overexpression improves tolerance to exogenous ethanol by bolstering the membrane integrity to either prevent membrane fluidization due to ethanol or to prevent diffusion of exogenous ethanol into the cell. In either case, these membrane modifications due to \textit{lpcA} overexpression may not improve ethanol production if the modified membrane likewise prevented endogenous ethanol from diffusing out of the cell. Alternatively, \textit{tilS} overexpression is directly involved in translation as opposed to cell membrane composition, which could similarly improve tolerance to both exogenous and endogenous ethanol. As shown in Figure 3-3B, we also found that ethanol production in these clones is growth-associated as all three clones show a similar correlation between the ethanol produced and the cell mass present. Considering that the \textit{Z. mobilis} homoethanol pathway uses pyruvate as the precursor to synthesize ethanol, it is not surprising that ethanol production is related to growth.

We recognized that another factor which could have a role in the discrepancy we observed here between tolerance and production is that the conditions under which we identified and verified the genes for ethanol tolerance differ in many aspects from the ethanol production system requirements. Some of the differing aspects include: aerobicity, ethanol shock, media/osmolarity, and strain genotype. The original tolerance testing conditions were microaerobic due to the low initial cell density and limited growth possible with the high ethanol stress (25 g/L) dosed initially, whereas the fermentations rapidly become anaerobic with higher inoculum cell densities and gradually accumulated ethanol. Achieving reasonable ethanol productivities of ethanol requires high cell densities (OD$_{600}$=5~10), and thus, we needed to switch from the MOPS minimal media with 2 g/L dextrose used for tolerance testing to a
production minimal media with higher salts concentrations. We used AM1 (5-fold greater ammonium and 20-fold greater phosphate concentrations) with 100 g/L dextrose, and in total, this increased the molarity of the media from about 0.1 M in the MOPS media to 0.6 M in the production media. Additionally, the genetic modifications in the production strain which prevent byproduct formation could alter the regulatory network in addition to the metabolic network.

To gain further insight into why *lpcA* and *tilS* did not improve ethanol production, we investigated whether these genes improve ethanol tolerance under production conditions by testing the effect of the *lpcA*- and *tilS*-overexpressing plasmids on ethanol tolerance (24-hr growth with 25 g/L ethanol dosing) in different growth conditions, specifically differing in the host genotype and media formulation. Growth during the first 24-hrs is important since the growth during the batch ethanol fermentations occurs almost entirely in the first 24 hours of fermentation and is related to the amount of ethanol produced as described above and depicted in Figure 3-3B. As shown in Figure 3-3C, the host strain and media have a pronounced effect on the trait-conferring quality of ethanol tolerance genes. Additionally and echoing the results of the batch fermentations shown in Figure 3-3A, *lpcA* and *tilS* do not confer improved growth under ethanol production conditions (LW05, AM1+IPTG), yet interestingly, the relative amount of initial growth between these clones is the same both in the presence of exogenously added high ethanol stress and with endogenous ethanol production, while the magnitude of growth in 24-hrs differs by approximately 6-fold for these two conditions. The *tilS* clone shows a 30 % reduction in the amount of growth in 24-hrs both in the presence of 25 g/L of exogenous ethanol and with endogenous ethanol production as compared to the growth of *lpcA* and empty vector control, which remain indistinguishable from one another in both conditions. These findings support that the growth environment largely affects whether and the extent to which genes confer
tolerance and further suggest that tolerance-conferring genotypes identified through directed genomics approaches will depend upon the growth conditions of the selective environment.

3.4 Conclusions

We have presented here an open-source ethanol production platform strain LW06, which has a fully characterized genotype and minimal genetic modifications that allow for ethanol production up to titers approaching the minimum inhibitory concentration and minimal fermentative byproducts production. This ethanol platform was designed to be genetically stable, have the ability to switch on ethanol biosynthesis, have a minimal number of pathway modifications, and thus provide a versatile platform for future genome engineering studies. To demonstrate the potential utility of a fully functional and debugged ethanol production chassis, we tested here how genes we previously identified to confer improved ethanol tolerance affect ethanol production and found that these genes did not improve ethanol production or ethanol tolerance under production conditions. In future work, this ethanol production system could be used to further improve ethanol production potentially using genome-wide approaches to search the genome and combinatorially optimize for desirable traits (Woodruff and Gill, 2011). For example, genome-wide selections could be designed to select for improved ethanol tolerance during production, which would also further investigate the tolerance-production relationship. Based on the discrepancies between ethanol tolerance and ethanol production that we identified here and attributed to the tolerance-conferring quality of genes being dependent on system conditions, future efforts to engineer product tolerance traits should be performed in a production system and under production-like conditions, which will not only help to directly compare studies but also elucidate the tolerance-production relationship more globally.
Building upon many ideas presented in the synthetic biology and metabolic engineering communities, we promote here the simple concept of a metabolic engineering commons for third-party distribution of platform production strains and have made a small step towards this goal by making our ethanol platform, which is not protected by any property rights by us or others to the best of our knowledge, available from the ATCC. Yet it’s important to also acknowledge that informal sharing of strains has been ongoing for a long time, but that there are limitations in some aspects of such sharing that a commons would be designed to get around. Moreover it is clear that there are additional design considerations that could be incorporated to yield better production chassis than presented here. For example, future efforts to construct platform strains with multiple deletions should consider using scarless deletion approaches (Blank et al., 2011; Sun et al., 2008) which would prevent introducing any new homology sequences throughout the genome (i.e., FRT-scars) which could allow for, though don’t usually lead to, deletions or inversions (Baba et al., 2006; Schweizer, 2003). Additionally, refactoring of inserted heterologous pathways to provide standardization and greater orthogonality would likely also be advantageous (Temme et al., 2012). We hope that the community will participate in a metabolic engineering commons since the usefulness of this open resource relies upon groups depositing strains, and indeed, many leaders in the field have expressed interest in doing so (personal communications from S.Y. Lee, J. C. Liao, J. Nielsen, D. S. Clark, C. A. Voigt, M. P. DeLisa, and G. N. Stephanopoulos).

3.5 Materials and Methods

3.5.1 Bacterial strains, plasmids, and media

Strains and plasmids used in this study are listed in Table 3-1. AM1 media without betaine supplementation (Martinez et al., 2007) and 100 g/L dextrose was used for fermentations and
tolerance testing as indicated. Tolerance testing was also performed using MOPS minimal media (Neidhardt et al., 1974) with 2 g/L dextrose. 1 mM IPTG was added to the medium to induce ethanol production where appropriate. Antibiotic concentrations used were: 100 μg/ml carbenicillin, 30 μg/ml kanamycin, 25 μg/ml chloramphenicol were added as appropriate. Aldehyde indicator plates were made freshly prior to use as previously described (Conway et al., 1987), and cells were plated on indicator plates and incubated at 37 °C for 16 hours in order for the color to develop. Primer synthesis and sequencing were performed by Operon EWG.

<table>
<thead>
<tr>
<th>Strain or plasmid</th>
<th>Relevant genotype</th>
<th>Reference or source</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW25113</td>
<td>$rrnB_{T14}$ $\Delta$ lacZ $\Delta$ hsdR514 $\Delta$ araBAD $\Delta$ rhaBAD $\Delta$BAD LD78</td>
<td>Datsenko and Wanner 2000</td>
</tr>
<tr>
<td>BW25113ΔrecA</td>
<td>BW25113, ΔrecA::FRT</td>
<td>Baba et al. 2006</td>
</tr>
<tr>
<td>LW01</td>
<td>BW25113, attTn7::pdc$<em>{zm}$ and adhB$</em>{zm}$ from Z. mobilis under lac promoter with Amp$^R$</td>
<td>This work</td>
</tr>
<tr>
<td>LW02</td>
<td>BW25113, attTn7::pdc$<em>{zm}$ and adhB$</em>{zm}$ from Z. mobilis under P$_{lacO1}$ Promoter with Amp$^R$</td>
<td>This work</td>
</tr>
<tr>
<td>LW03</td>
<td>LW02, ΔadhE::FRT</td>
<td>This work</td>
</tr>
<tr>
<td>LW04</td>
<td>LW02, ΔadhE, ΔackA::FRT</td>
<td>This work</td>
</tr>
<tr>
<td>LW05</td>
<td>LW02, ΔadhE, ΔackA, ΔfrdABCD::FRT</td>
<td>This work</td>
</tr>
<tr>
<td>LW06</td>
<td>LW02, ΔadhE, ΔackA, ΔfrdABCD, ΔldhA::FRT</td>
<td>This work (ATCC BAA-2466)</td>
</tr>
<tr>
<td>pLOI297</td>
<td>pUC18 backbone containing pdc$<em>{zm}$ and adhB$</em>{zm}$ from Z. mobilis under lac promoter with Amp$^R$ and Tet$^R$</td>
<td>ATCC 68239 (Alterthum and Ingram 1989)</td>
</tr>
<tr>
<td>pSMART-LCKan</td>
<td>Cloning vector with Kan$^R$</td>
<td>Lucigen</td>
</tr>
<tr>
<td>tilS-pSM</td>
<td>pSMART-LCKan vector containing tilS</td>
<td>Woodruff et al., in press</td>
</tr>
<tr>
<td>lpcA-pSM</td>
<td>pSMART-LCKan vector containing lpcA</td>
<td>Woodruff et al., in press</td>
</tr>
</tbody>
</table>

**Table 3-1** Strains and plasmids used in this study.
3.5.2 Genetic methods

Homologous recombination for chromosomal insertions was performed using pSIM5 which contains temperature-inducible expression of the λ bacteriophage recombination proteins gam bet exo (Datta et al., 2006). Strains were transformed with pSIM5 and selected for chloramphicol resistance at 30 °C. Recombineering was performed as previously detailed (Sharan et al., 2009). Briefly, 15 ml of SOB in a 125 ml baffled flask was inoculated with 150 μl of an LB overnight culture (30 °C) and incubated at 30 °C until the culture reached an OD₆₀₀ = 0.4-0.6. The culture was induced in a 42 °C water bath for 15 min with continuous shaking, and subsequently chilled in an ice-water bath for 5 min with frequent mixing. The culture was transferred to a 50 ml conical tube and centrifuged at 4 °C and 5,000 rpm for 4 min. The pellet was washed twice with 15 ml of ice-cold ultra-pure water and finally resuspended in approximately 150 μl of ultra-pure water. 80 μl of resuspended cells and 1 μl (200 ng) of dsDNA insert containing 50 bp homology arms (gel-purified and concentrated by ethanol precipitation) were added to a chilled 1 mm gap electroporation cuvette (Eppendorf), mixed by pipetting up and down, and electroporated at 1800 V (Eppendorf).

FRT-flanked kanamycin resistance markers were removed using pCP20, which contains the temperature-inducible Flippase recombination enzyme (Flp recombinase) from S. cerevisiae, as previously described (Cherepanov and Wackernagel, 1995; Datsenko and Wanner, 2000). All primers and PCR reactions used for strain constructions are listed in Table 3-2. Primers used for gene deletions were based on those used for the Keio collection construction (Baba et al., 2006). All primers used for verification by strain modifications by sequencing are listed in Table 3-3. pSIM5 and pCP20 were removed by serial growth at 37 °C and 42 °C, respectively. Standard protocols for PCR amplification and ethanol precipitation were used (Sambrook and Russell,
2001). Gel purification was performed using QIAquick gel extraction kits (Qiagen), and plasmids were purified using QIAprep Miniprep Spin Kits (Qiagen). Genomic DNA was purified using a PureLink Genomic DNA Kit (Invitrogen).
<table>
<thead>
<tr>
<th>Genetic Target</th>
<th>Template</th>
<th>Left nt</th>
<th>Right nt</th>
<th>Left primer (5’→3’)</th>
<th>Right primer (5’→3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{lac}$-pdc-$adhB$ (left 50bp H1 attTn7 homology)</td>
<td>pLOI297</td>
<td>3909746</td>
<td>3909795</td>
<td>tctgtaacagaaaaagtagccgaagatgagttgctcaactggttaacctctaggccgacc</td>
<td>caagacagggacggaaaaaccg</td>
</tr>
<tr>
<td>$Amp^R$ cassette (left 20bp $adhB$ overlap, right 50 bp attTn7 H2)</td>
<td>pLOI297</td>
<td>3909845</td>
<td>3909797</td>
<td>cgggttctgcctgtctgcgcggaaaccctattttttt</td>
<td>gtaagcgggcaatttttctcgtatgtttataccaaacatctctcgctgaaggttcgacc</td>
</tr>
<tr>
<td>attTn7 insertion cassette amplify (zip ETOH operon &amp; $Amp^R$)</td>
<td>$P_{lac}$-$adhB$ &amp; $Amp^R$</td>
<td>3909795</td>
<td>3909797</td>
<td>tctgtaacagaaaaagtagccgaagatgagttgctcaactggttaacctctaggccgacc</td>
<td>gtaagcgggcaatttttctcgtatgtttataccaaacatctctcgctgaaggttcgacc</td>
</tr>
</tbody>
</table>
| Amplify $P_{UacO-1}$.75 bp (overlapping primers)        | none      | n/a     | n/a      | Ataaaatgtgagcggatatggactgtgtagcggacggagaagaaga                                      | ataagctgtagcgatgctggtctgctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggctggc

Table 3-2 PCR reactions and primers used for strain construction
<table>
<thead>
<tr>
<th>Genetic Target</th>
<th>Left nt</th>
<th>Right nt</th>
<th>Left primer (5'→3')</th>
<th>Right primer (5'→3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>attTn7 EtOH insert verify</td>
<td>3909511</td>
<td>3909929</td>
<td>GCGACAACAGTTGCGACGGT</td>
<td>TCGCGCTGATCAAAGGCACC</td>
</tr>
<tr>
<td>ΔadhE verify</td>
<td>1294609</td>
<td>1297425</td>
<td>TCC TTA ACT GAT CGG CAT TG</td>
<td>CAAATCATCACCGCACTGAC</td>
</tr>
<tr>
<td>ΔfrdABCD verify</td>
<td>4376905</td>
<td>4380589</td>
<td>AGGGGCAGCAAATGTGGAGCA</td>
<td>CAAACAGCGGTGGCGAGTGACT</td>
</tr>
<tr>
<td>ΔackA verify</td>
<td>2411259</td>
<td>2412900</td>
<td>TCCTTAGCAGCCTGAAGGCT</td>
<td>ACGGGCTGAGCGATAGGTTT</td>
</tr>
<tr>
<td>ΔldhA verify</td>
<td>1439719</td>
<td>1441147</td>
<td>ATCCCGCCATCAGCAGGCT</td>
<td>TCATCAGCAGGTCAACCGCA</td>
</tr>
</tbody>
</table>

Table 3-3 Primers used for strain verification

At each construction step, strains were serially cultured and genotyped by gel electrophoresis and sequencing to ensure genetic homogeneity. Each cell contains approximately 4-8 copies of the chromosome, and with the efficiency of dsDNA recombination being approximately 0.1%, a single recombinant cell probably contains only one copy of the desired recombinant chromosome initially. Therefore, positive transformants must be serially selectively cultured (typically ~5 serial platings) to allow for chromosomal segregation and homogeneity of the desired recombinant chromosome.

3.5.3 Fermentations and tolerance testing

Batch fermentations were inoculated to \( \text{OD}_{600} = 0.2 \) in 1 L Biostat A-plus bioreactors (Sartorius) with 0.5 L final volume of AM1 media (100 g/L dextrose) with IPTG and carbenicillin (kanamycin addition if plasmids present). Fermentors were maintained at 37 °C with 400 rpm mixing, 4 °C exhaust condenser, and pH 7 controlled by automatic addition of 10 M KOH. For the inoculum starter culture, overnight 15ml cultures started from freezerstocks were grown in AM1 media with 100 g/L dextrose and antibiotics in 50 conical tubes, and the next day, 100 ml of AM1 media (100 g/L dextrose) with antibiotic was inoculated from the
overnight culture to an \( \text{OD}_{600} = 0.1 \) in a 500 ml baffled flask. Starter flasks were grown to an \( \text{OD}_{600} = 0.5 \) at 37 °C and then used to inoculate fermentors. Fermentations were performed in duplicate.

Tolerance tests were performed in AM1 media with 100 g/L dextrose and MOPS minimal media with 2 g/L dextrose. IPTG (1 mM) was added only to the media dosed with 25 g/L ethanol as indicated and was not added to starter cultures. 5 ml overnight cultures in 15 ml conical tubes were grown in AM1 or MOPS media (depending on which media was used in tolerance test). The overnight cultures were used the next day to inoculate 5 ml starter cultures, also in AM1 or MOPS media (depending on which was used in tolerance test), to an \( \text{OD}_{600} = 0.2 \) and grown at 37 °C to \( \text{OD}_{600} = 0.4 \). Tolerance tests were inoculated at an \( \text{OD}_{600} = 0.04 \) in 5 ml total volume of AM1 or MOPS media and 25 g/L ethanol dosed in 15-ml screw-cap conical tubes and maintained in a shaking incubator (37 °C, 225 rpm). Growth (\( \text{OD}_{600} \)) was measured after 24 hr of incubation.

3.5.4 Analytical techniques

Using a UV-Vis spectrophotometer (Shimadzu), growth was monitored by the \( A_{600} \) (\( \text{OD}_{600} \)), and DNA was quantified using the \( A_{260} \). Ethanol and glucose in the fermentation supernatants were quantified using a 2700 Bioanalyzer (YSI) and the manufacturer’s protocols. Organic acids in the fermentation supernatant samples were quantified by HPLC (Shimadzu) using an Aminex HPX-87H column at 60 °C (0.6 ml/min 0.005 M \( \text{H}_2\text{SO}_4 \)) and a refractive index detector (Shimadzu). Organic acids were also verified using metabolite-specific enzymatic kits (R-Biopharm).
Acknowledgements

We thank ML Lipscomb for advice regarding fermentations and NR Boyle for critical reading of an earlier version of this manuscript. LBA Woodruff was supported by a NSF Graduate Research Supplement (CBET0449183).

3.6 References

Arkin AP, Fletcher DA. 2006. Fast, cheap and somewhat in control. Genome Biology 7(8).


Chapter 4 Engineering improved ethanol production in

*Escherichia coli* with a genome-wide approach

Accepted pending revision at *Metabolic Engineering* (November 2012).

4.1 Abstract

A key challenge to the commercial production of commodity chemical and fuels is the toxicity of such molecules to the microbial host. While a number of studies have attempted to engineer improved tolerance for such compounds, the majority of these studies have been performed in wild-type host strains and laboratory culturing conditions that differ considerably from what will be employed for production. Here we have applied the multiscalar analysis of library enrichments (SCALEs) method and performed a growth selection in an ethanol production system to quantitatively map in parallel all genes in the genome a metabolically engineered E. coli ethanol production strain onto ethanol tolerance and production. We compared the identified genes to prior efforts using a wild-type strain and laboratory conditions, and then engineered a new strain with improved ethanol tolerance and productivity. In order to perform the selection in an ethanol-producing system, we used the previously constructed metabolically engineered E. coli LW06 strain (ATCC BAA-2466), which produces ethanol (>30 g/L, 80% theoretical yield in minimal media) by heterologous expression of the Zymomonas mobilis pathway for ethanol biosynthesis (Woodruff et al, submitted), ethanol production platform was used as the host strain for the constructed multiscalar genomic libraries (>106 clones for each library of 1, 2, or 4 kb overlapping genomic fragments at 125 nt resolution intervals on average with 106 clones) in this production strain. By testing individually selected clones, we confirmed that the growth selections were employed to enriched for clones
with both improved ethanol tolerance and production phenotypes. After confirmation of
tolerance and production phenotypes, we performed combinatorial testing of the top genes
identified (uspC, otsA, otsB) to investigate their ability to confer improved ethanol tolerance or
ethanol production. We determined that overexpression of otsA was required for improved
tolerance and productivity phenotypes, with the best performing strains showing up to 75%
 improvement relative to the previously described production strain.

4.2 Introduction

Metabolic engineering has been used to produce a wide range of chemicals (Bailey, 1991;
Keasling, 2008; Stephanopoulos et al., 1998). Many of these chemicals, as well as the
intermediates involved in their biological production, inhibit growth and/or metabolism, thus
reducing overall performance of the engineered strain (Dunlop, 2011; Nicolaou et al., 2010;
Warnecke and Gill, 2005). Tolerance engineering often entails rounds of random mutagenesis,
typically by adaptive- or chemically-induced mutagenesis, and growth selections (Atsumi et al.,
2010; Horinouchi et al., 2010; Minty et al., 2011; Smith and Liao, 2011; Wang et al., 2008;
Yomano et al., 1998). Advances in DNA sequencing have recently enabled whole-genome
sequencing of the selected mutants. Identified mutations are then reconstructed one at a time in
the original host strain to identify trait-conferring mutations (Atsumi et al., 2010; Minty et al.,
2011). While this process represents a considerable advance from prior capabilities, isolating the
trait-conferring mutation(s) from all of the other randomly acquired mutations remains laborious
since beneficial mutations make up only a small fraction of the total (5 out of all 28 mutations
tested in (Atsumi et al., 2010), 0.01-1% for a protein (Romero and Arnold, 2009)). Several
directed genome-engineering approaches, which can systematically mutate and track genes at the
genome-scale, were previously developed to address this challenge. Such approaches have been
used to rapidly map functional genotype-phenotype relationships throughout the entire genome and engineer a range of improved traits (Baba et al., 2006; Borden and Papoutsakis, 2007; Cho et al., 1998; Gill et al., 2002; Goodarzi et al., 2010; Lynch et al., 2007; Smith et al., 1995; Warner et al., 2010b). Here we have tailored a directed, genomic approach to map the *E. coli* genome for improved ethanol tolerance and ethanol production using an *E. coli* ethanol production strain.

The production of alcohol biofuel is currently limited by the highly inhibitory nature of short-chain alcohols (C$_2$-C$_8$) (Ingram, 1990; Ingram and Buttke, 1984). Therefore, various efforts to engineer improved product tolerance have been undertaken in order to improve production, yet only some studies have found that improved tolerance translates into improved production (Dunlop, 2011). In our previous work, we identified ethanol tolerance genes in a wild-type host (Woodruff et al., in press) and determined that the tolerance-conferring quality of identified ethanol tolerance genes was largely dependent on aspects of the growth environment (*e.g.*, media, osmolarity) and microbial host (*e.g.*, strain genotype, metabolism). As a result, ethanol tolerance genes identified in a wild-type host strain did not improve ethanol production in the metabolically engineered production strain (Woodruff et al., submitted). Here we employed a similar strategy using the SCALEs method but now using an *E. coli* production strain engineered to produce ethanol at levels similar to those previously published (approx. 40 g/L) (Alterthum and Ingram, 1989; Jarboe et al., 2007; Ohta et al., 1991; Yomano et al., 2008).

Previous studies have focused on improving ethanol production and tolerance in *E. coli*. Several have used adaptation and screening (typically adapting over 50-90 days) to identify mutated strains with improved tolerance and production of ethanol (Wang et al., 2008; Yomano et al., 1998; Yomano et al., 2008). Additionally, *E. coli* ethanol production and tolerance has been improved through expression and random mutagenesis of the exogenous global regulator
IrrE (Chen et al., 2011b; Ma et al., 2011), which is based upon the global transcription machinery engineering approach (gTME) developed by Alper, et al. (2006) that was also used to improve ethanol tolerance in *E. coli* (Alper and Stephanopoulos, 2007). Adaptation has also been used to improve tolerance to other alcohols in *E. coli*. Adaptation to isobutanol resulted in increased tolerance to but not production of isobutanol, (Atsumi et al., 2010) and iterative rounds of chemically-induced, whole-cell random mutagenesis coupled with screening for tolerance to amino acid-analogues (anti-metabolites) and production of the alcohol improved the production of 3-methyl-1-butanol (Connor et al., 2010) and isobutanol (Smith and Liao, 2011).

In this work, we demonstrate the application of a directed, genome-wide approach to improve ethanol biofuel production. By using our previously constructed ethanol production platform strain (LW06), designing selections for ethanol tolerance under ethanol production conditions, and applying the multiSCalar Analysis of Library Enrichments (SCALEs) genome-wide approach (Lynch et al., 2007), we mapped the genome for ethanol production and identified clones conferring improved ethanol tolerance and ethanol production. We performed further combinatorial analysis to identify the overexpressed genes responsible for the improved phenotype.

4.3 Materials and methods

4.3.1 Strains, media and reagents

The LW06 *E. coli* ethanol production strain (BW25113 ΔldhA ΔackA ΔfrdABCD ΔadhE attTn7::P<sub>LlacO-1</sub> pdc<sub>Zm</sub> adhB<sub>Zm</sub> Amp<sup>R</sup>), available from the ATCC (# BAA-2466), was used for all experiments (Woodruff et al, submitted). A modified version of the AM1 minimal media previously optimized for ethanol and lactate production (Martinez et al., 2007), which we termed AMX media, was used for the selections, fermentations, and growth testing. AMX media does
not contain betaine supplementation and additionally contains 0.1 M MOPS buffer (1M stock solution buffered to pH 7.2 with 10 M KOH) and 0.1 M monobasic potassium phosphate (1M stock solution buffered to pH 7.4 with 10 M KOH) for pH buffering but is otherwise identical to the AM1 media recipe. Buffers and concentrations were experimentally determined to maximize batch ethanol production of LW06 at 100 hr with a final pH ≥ 6. AMX media was freshly prepared prior to use. Carbenicillin (100 μg/ml) was added to all LW06 cultures, and kanamycin (30 μg/ml) was added to all cultures containing the pSMART-LCKan vector. Molecular biology grade ethanol was used for ethanol dosing (Sigma-Aldrich E7148). Primer synthesis and sequencing was performed by Operon.

4.3.2 Library construction

Genomic DNA was extracted from 100 ml of an overnight culture of LW06 in LB medium using a PureLink Genomic DNA kit (Invitrogen). The genomic DNA was concentrated to 2 μg/μl by ethanol precipitation (Sambrook and Russell, 2001). Digestions reactions (50 μl) containing 50 μg aliquots of the purified genomic DNA (A$_{260}$/A$_{280}$ > 1.8) were digested with 2.5 U each of RsaI and AluI (NEB) in NEBuffer 4 for 0.5, 1, 2, 3, 5, and 10 min at 37 °C followed by inactivation at 65 °C for 20 min. RsaI and AluI are blunt-cutting restriction endonucleases that cleave 4 bp recognition sequences and are used to randomly digest genomic DNA at 128 bp intervals on average (every 4$^4$bp per enzyme / 2 enzymes). Digestion times and enzyme loading were determined experimentally from trial digestions to provide approximately uniform distributions of genomic fragments ranging 0.1 – 12 kb in size. All six digestion reactions were mixed and separated by gel electrophoresis. Fragments for each library of approximately 1 kb (0.8 – 1.2 kb), 2 kb (1.8 – 2.2 kb), and 4 kb (3.5 – 4.5 kb) were excised and purified using a QIAquick gel extraction kit (Qiagen). The 1, 2, and 4 kb genomic libraries were constructed in
the pSMART-LCKan vector (Lucigen) according to manufacturer’s protocol except ligation reactions were additionally incubated at 4 °C overnight. Before transformation, each 10 μl library ligation reaction was concentrated by standard ethanol precipitation with yeast tRNA (pellets dissolved in 2 μl nuclease-free water). Each library transformation was performed at 1800 V with 60 μl of *E. coli* 10G elite electrocompetent cells (Lucigen) and 2 μl of the concentrated ligation reaction and recovered in 1 ml SOC for 1 hr before dilution plating to ensure > 3x10^5 total transformants per library which is 99.9+ % probability of entire genome representation at 125 bp intervals from the equation: \( N = \ln(1-P)/\ln(1-f) \), where \( f \) is the fraction of the genome (125 bp/4.64x10^6 bp) (Clarke and Carbon, 1976). The remaining library recovery was mixed with 15 % glycerol (v/v) and frozen at -80 °C until plating. Each genomic library was plated on LB plates containing kanamycin with approximately 10^4 CFU/plate and incubated at 37 °C overnight before all plates for each library were scraped with a cell lifter. Plasmid DNA from the cell pellet for each library was extracted using a QIAprep Spin Miniprep kit (Qiagen). The 1, 2, and 4 kb plasmid libraries were transformed into LW06 made electrocompetent by washing with ice-cold ultrapure water. Electroporations for each library were performed in duplicate with each cuvette containing about 10^10 cells and 200 ng of plasmid DNA (>10^6 transformants per library in LW06). Each library was amplified by plating 10^6+ total transformants (~5x10^4 CFU/plate) on LB plates with kanamycin, and plates were incubated at 37 °C for 15 hr before scraping all plates for each library into LB. Plasmid DNA from an aliquot of each library was extracted, digested with SwaI, and verified by gel electrophoresis to contain genomic fragments of the correct size. The genomic libraries of LW06 contained in the LW06 host were stored at -80 °C in 15 % glycerol (v/v) with each tube containing ~3x10^10 cells.

4.3.3 Selections and microarray quantification
One frozen tube of each of the 1, 2, and 4 kb libraries was thawed and added to 50 ml of AMX media with 10 g/L glucose pre-warmed to 37 °C (10^{11} cells total), and this culture was used to inoculate 100 ml of AMX (10 g/L glucose, antibiotics) to an OD of 0.2 in a 250 ml Erlenmeyer flask. The starter culture was grown to an OD = 0.4, then used to inoculate the selection to an OD of 0.04, and samples for later microanalysis were frozen. Selections were performed in 125 ml Erlenmeyer flasks stoppered with a serum stopper (to prevent ethanol evaporation) vented with a 22 G needle, and each flask contained a total volume of 50 ml of AMX media containing antibiotics, 100 g/L glucose, 1mM IPTG (to induce ethanol production), and ethanol dosed to the desired final concentration. Selections with 15 and 30 g/L of exogenous ethanol were inoculated, but the 30 g/L selection did not grow after the first batch for either the genomic libraries or the adaptation (>1 month). Adaptation of LW06 containing the empty vector was identically performed except starting from a frozen stock of this strain containing approximately 10^{11} cells. Subsequent batches were inoculated identically every 24 hours for a total of 8 batches. Selections were incubated at 37 °C with shaking. Samples of the selection cultures were pelleted and stored frozen in 15 % glycerol (v/v) at -80 °C until analysis. The final batch of the genomic libraries selection and adaptation were plated to select individual colonies for testing and sequencing; one colony was randomly selected from the adapted culture.

Microarray analysis was performed on the initial mixed library culture and the selected culture from the end of the fourth and eighth batches (15 g/L selection). Samples were prepared by extracting the plasmid DNA from approximately 10^{10} cells for each time point using a QIAprep Spin Miniprep kit (Qiagen). Control DNA (pGIBS-DAP, -THR, -TRP, -PHE) was prepared by extracting plasmid DNA from an overnight culture grown in LB with 50 μg/ml carbenicillin. Plasmid DNA samples were concentrated to 0.4 μg/μl by standard ethanol
precipitation. The DNA sample for each microarray was prepared by mixing 3 μg of selection plasmid DNA (initial, batch 4, or batch 8) with control DNA (3 μg pGIBS-DAP, 12 ng pGIBS-THR, 6 ng pGIBS-TRP, 3 ng pGIBS-PHE) and fully digested with 10 U each of Rsal and AluI (NEB) in NEBuffer 4 at 37 °C overnight before heat activation at 70 °C for 15 minutes. A 5.6 μl aliquot of One-Phor-All buffer stock (10X) (discontinued, prepared according to GE Healthcare recipe: 100 mM Tris-acetate at pH 7.5, 100 mM magnesium acetate, 500 mM potassium acetate) and 1 U DNase I (NEB) was added to each reaction and incubated at 37 °C for 5 minutes, followed by inactivation at 98 °C for 20 minutes. A prior titration assay was performed to determine the DNase I dosage to yield fragments of 50-100 bp on average according to the microarray manual (Affymetrix). 100 U of Exonuclease III (NEB) were added and incubated at 37 °C for 15 then 98 °C for 20 minutes followed by 3’ termini labeling using a BioArray Terminal labeling kit with biotin-ddUTP (Enzo) according to the manufacturer’s protocol.

Samples were hybridized to GeneChip E. coli Antisense Genome Arrays (Affymetrix), washed, stained, and scanned at the University of Colorado Microarray facility according to the manufacturer’s protocols. The microarray signals were decomposed into clonal fitness values \( \frac{freq_{\text{final}}}{freq_{\text{initial}}} \) for detected genomic fragments using the SCALEs software (Lynch et al., 2007) as previously described (Warnecke et al., 2008). The fitness for each gene was calculated as the sum of clonal fitness values for all clones reported to contain any part of the gene.

4.3.4 Ethanol tolerance testing

Overnight cultures and inoculum for tolerance tests were grown in AMX media with 10 g/L glucose and antibiotics at 37 °C with shaking. A 1 ml overnight culture in a sterile 1.5 ml microcentrifuge tube was inoculated from a toothpick transfer of a frozen stock of the clone. Overnight cultures were used to prepare mid-exponential inoculum by diluting the culture to
OD\textsubscript{600} = 0.1 in 6 ml total in a 15 ml conical tube and growing the culture until it reached an OD\textsubscript{600} = 0.4. Tolerance tests were inoculated at an OD\textsubscript{600} = 0.04 in AMX media with 100 g/L glucose, 1 mM IPTG, antibiotics, and ethanol dosed to the desired final concentration. Growth after 24 hr was measured using a UV-Vis spectrophotometer (Shimadzu) at 600 nm. All tolerance tests were performed in triplicate starting from separate overnight cultures. P-values were calculated from the triplicate OD\textsubscript{600} measurements after 24 hr of growth using a two-tailed, independent $t$-test (assuming unequal variance as represented in the data).

4.3.5 Batch fermentations

Overnight and inoculum cultures were prepared identically as those prepared for tolerance testing. Batch fermentations were carried out in 125 ml Erlenmeyer flasks containing a total culture volume of 30 ml at an initial OD\textsubscript{600} = 0.04 in AMX media with 100 g/L glucose, 1 mM IPTG, and antibiotics. Flasks were stoppered with rubber septa to prevent ethanol evaporation and vented with a 22 G needle. A 1 ml sample was removed initially and every 24 hr for OD\textsubscript{600} measurement and ethanol quantification from the supernatant. Ethanol was quantified enzymatically using a 2700 Bioanalyzer (YSI) and by HPLC with an HPX-87H column (Aminex) at 60 °C (0.6 ml/min 0.005 M H\textsubscript{2}SO\textsubscript{4}) and a refractive index detector (Shimadzu). All fermentations were performed in triplicate starting from separate overnight cultures, and p-values were calculated for the triplicates using a two-tailed, independent $t$-test (unequal variance).

4.3.6 Clone construction

The following primers were designed to amplify the genes of interest from E. coli MG1655 genomic DNA (ATCC 47076): \textit{uspC}\_1977487\_f (5′–acagacacttaagcgcaacaact–3′), \textit{uspC}\_1978255\_r (5′–ccccagcgcgcgataaggctc–3′), \textit{otsA}\_1978130\_f (5′–agcatctgtgccggttga–3′), \textit{otsA}\_1979711\_r (5′–cggtgtgctggctggc–3′), \textit{otsB}\_1979585\_f (5′–
cgtctcgctggtctggtgcaat, otsB_1980503_r (5′–aaatggcaccccgctaca–3′). These primers were phosphorylated and used to amplify the desired genomic region using Platinum Pfx DNA polymerase (Invitrogen) according to the manufacturer’s recommendations. PCR products were purified by electrophoresis using a QIAquick Gel Extraction kit (Qiagen) and ligated to the pSMART-LCKan backbone using the CloneSMART kit (Lucigen) with incubation overnight at 4 °C and 2 hr at 25 °C. Following heat inactivation at 70 °C for 20 min, 1 μl of each ligation reaction was transformed into 80 μl of LW06 made electrocomponent by ultra-pure water washing. Clones were selected on LB plates with kanamycin. Plasmid sizes were verified by gel electrophoresis following digestion with SwaI to check for chimeras and plasmid insert sequences were verified by sequencing with primers SL1 (5′–cagtccagttacgctggagtc–3′) and SR2 (5′–ggtcaggtatgatttaaatggtcagt–3′).

4.4 Results and Discussion

4.4.1 Designing a directed, genome-wide selection for ethanol production

We used the SCALEs approach to perform a high-resolution genome-wide search for genetic targets that confer improved ethanol tolerance during ethanol production. The SCALEs approach combines i) multiple plasmid based genomic libraries with each library containing genomic fragments of a defined size (1, 2, or 4 kb), ii) DNA microarray analysis of plasmid DNA throughout selections, and iii) a multiscalar decomposition algorithm used to map DNA microarray data back onto individual clones in the selection (Gill et al., 2002; Lynch et al., 2007). We previously constructed an E. coli ethanol production platform, LW06 (Figure 4-1A) and using this platform, found that ethanol production was correlated to growth in multiple genetically distinct clones, which was expected given that the ethanol production pathway diverts pyruvate to ethanol biosynthesis (Woodruff et al, submitted). Therefore, we
hypothesized here that selecting for improved growth in ethanol production conditions would similarly select for improved ethanol production.

To investigate this overarching hypothesis, we constructed SCALEs genomic libraries containing 1, 2, or 4 kb fragments of the genomic DNA from the LW06 (BW25113 ΔldhA ΔackA ΔfrdABCD ΔadhE attTn7::PₐlacO₁ pdcZₘ adhBZₘ AmpR) production strain in a low-copy
(~20 copies/cell) vector (each library with >99.9% probability of genome representation at 125 bp intervals based on the yield of transformants). Notably, a previous study using SCALEs to improve succinate production found that it was necessary to use the genomic DNA from the host for library construction to prevent complementation by genes disrupted in the production strain (Singh et al., 2009). Each plasmid library was then transformed into the LW06 host so that the selections could be performed under conditions of intracellular ethanol stress experienced during ethanol production (Figure 4-1B).

The selective environment was designed to investigate ethanol tolerance under laboratory production conditions, which attempt to approximate the complex production environment found in industrial bioreactors (e.g., high osmolarity, endogenously produced ethanol, and high exogenous ethanol concentrations). Specifically, selections were performed upon an equimolar mixture of the three genomic libraries (1, 2, and 4 kb insert sizes) in production media (AMX with 100 g/L glucose and 1 mM IPTG) and 15 g/L of exogenous ethanol added (Figure 4-1C). The selection was serially transferred every 24 hr into identical conditions for a total of 8 serial batches. During batch fermentations, growth typically stalls when the supernatant has accumulated 15-20 g/L ethanol (~48 hrs). After 8 batches of growth, the selection culture showed a definitive trend of increased growth per batch (Figure 4-1D) indicating that we had enriched the population for clones with improved growth under these conditions. As a comparison, we performed in parallel an identical selection upon the LW06 parent strain containing the empty vector. However, growth of this culture did not show any improvement, as should be expected given the time required to generate relevant diversity by random adaptation alone (Figure 4-1D). We also performed a selection at 30 g/L of ethanol, but there was no growth of this culture after the first serial transfer (1 month incubation) suggesting that 30 g/L
was too strong of a selective pressure when added on top of intracellular ethanol production (since in prior studies 30 g/L in a wild-type \textit{E. coli} host were successful (Woodruff et al, in press)). Figure 4-1D also shows the interesting trend (which we have observed in all of our ethanol selections) that the growth is substantially greater in the first batch than in the subsequent batch(es), which suggests that the initial ethanol stress may not be as inhibitory as repeated ethanol exposure.

\textbf{4.4.2 \textit{E. coli} genome mapping for ethanol production trait-conferring genotypes}

To quantify genome-wide enrichment patterns throughout selection, samples of the plasmid DNA were isolated from the initial mixed library, the fourth batch (16.5 generations), and at the end of the eighth batch (33 generations) and analyzed by genome microarrays (Figure 4-1C). Using the SCALEs algorithm (Lynch et al., 2007), the fitness values (W= frequency$_{\text{final}}$/frequency$_{\text{initial}}$) for all $\sim$13,000 clones ($10^{10}$ cells/sample) present at the two selection time points were calculated (Figure 4-2A, Figure 4-2B). As shown in Figure 4-2A, many genomic regions were differentially enriched at batch 4, and interestingly, many of these genes encode for products involved in biological processes previously linked to ethanol stress and tolerance. For example, \textit{fabA} (gene rank=9), which introduces the double bond in unsaturated fatty acid biosynthesis (UFA) and has complex regulation of transcription (Feng and Cronan, 2009), has previously been shown to improve ethanol tolerance when it is overexpressed (Luo et al., 2009). MrdA (gene rank=2) is a peptidoglycan synthetase with transpeptidase activity required for cross-linking of peptidoglycan and maintains the cell’s rod shape against osmotic pressure (Ishino et al., 1986; Sauvage et al., 2008). Ethanol has previously been shown to specifically inhibit cross-linking of peptidoglycan and lead to cell lysis (Ingram and Buttke, 1984; Ingram and Vreeland, 1980). We also found enrichment of many genes involved in acid
resistance (hdeA, hdeB, dctR, yhiD), which is contradictory to previous genome mapping for ethanol tolerance (Goodarzi et al., 2010), but similar to previous studies on isobutanol stress in which genes involved in acid resistance exhibited increased expression, including hdeA and hdeB (Brynildsen and Liao, 2009).

Figure 4-2 Genome enrichment profiles during ethanol selection. Chromosomal maps showing the clonal fitness (final concentration over the initial concentration for clone i) for all quantified clones at the fourth batch (A) and eighth batch (B) of selection. Each peak on the chromosomal represents a clone; the height of the peak corresponds to the clonal fitness; the color (red=1 kb, yellow=2 kb, green=4 kb) denotes the size of the clone’s genomic fragment; and the location around the chromosome (genome nt 1 at 12 o’clock) corresponds to the locus of the genomic fragment. (C) The fitness for each gene at each quantified selection time point was quantified by summing the clonal fitness values for all clones containing the gene. The locus corresponding to the uspC, otsA, and otsB genes was most highly enriched at the end of the ethanol production selection.
A number of genes (\textit{dctR, dksA, xylR, sfsA}) are also involved in transcription regulation, which has previously been shown to be involved in eliciting ethanol tolerance (Woodruff et al, in press) (Alper and Stephanopoulos, 2007). FdnG and UbiB are believed to be involved in electron transfer and maintenance of the proton motive force (PMF). Ethanol is known to dissipate the PMF (Sikkema et al., 1995), and previously ethanol stress was found to upregulate the \textit{psp} operon thought to restore the PMF (Brynildsen and Liao, 2009). Some of the other biological processes that these highest fitness clones are involved in are: transport (\textit{yhiD, tatA, tatB}), DNA repair (\textit{mutL, dksA}), and translation (\textit{rlmH, truB, gluQ}).

In contrast to the relatively high genetic diversity of clones enriched at the fourth batch, by the final batch the culture is predominantly enriched with a single genomic region containing the three genes \textit{uspC, otsA}, and \textit{otsB} (Figure 4-2B). We calculated the gene fitness values for all genes as the sum of clonal fitness values for all clones containing any region of the gene. While many of the highest fitness genes have nearly equivalent values at the fourth batch, the eighth batch has much greater separation of the highest fitness values (Figure 4-2C), which is expected with more generations of growth to differentiate the frequencies of clones with similar growth rates. The genome enrichment pattern here was also found to differ substantially from previous ethanol tolerance selections performed in a wild-type host strain (Woodruff et al, in press) using a similar SCALEs approach (Figure 4-3).
Figure 4-3 Comparison of genome-wide enrichment in the production selection approach presented here to previously performed ethanol tolerance selections in a wild-type host strain (Woodruff et al., in press) using a similar SCALEs strategy. Previous tolerance selections were performed with 15 and 30 g/L of exogenous ethanol. All gene fitness values were calculated as described in this work.

The uspC-otsA-otsB genomic region encodes for a universal stress protein (uspC) and the two enzymes that comprise the trehalose biosynthesis pathway (otsA, otsB). UspC has been shown to be involved in resistance to UV-induced DNA damage and act as a scaffold during osmotic stress for the KdpDE two-component system that regulates intracellular K⁺ during osmotic shock (Gustavsson et al., 2002; Heermann et al., 2009). Trehalose is a nonreducing disaccharide that is widely used in nature and a variety of applications to stabilize membranes or proteins and preserve desiccated cells; it prevents protein denaturation and aggregation, and the mechanism is believed to be due to the high hydrophilicity, stability, and large hydrated radius of trehalose (Crowe et al., 1984; Crowe et al., 2001; Jain and Roy, 2009; Sola-Penna and Meyer-Fernandes, 1998; Xie and Timasheff, 1997). In vivo trehalose is accumulated in response to a
variety of stressors and improves resistance to these stresses, such as heat shock (Hottiger et al., 1994; Piper, 1993), cold shock (Kandror et al., 2002), desiccation (Crowe et al., 1984; de Castro et al., 2000; Miller and Ingram, 2008; Pilon-Smits et al., 1998), oxidative stress (Benaroudj et al., 2001), and osmotic stress (Styrvold and Strom, 1991), in a variety of organisms (Elbein et al., 2003). Additionally, trehalose has been found to improve ethanol tolerance and ethanol production in yeast (Mansure et al., 1994; Pham and Wright, 2008; Soto et al., 1999; Wu et al., 2006). In response to osmotic stress, *E. coli* synthesizes and accumulates trehalose and glutamate, which provide moderate osmotic tolerance, while supplementation with betaine, proline, or choline can provide a higher degree of osmotic tolerance (Larsen et al., 1987; Strom et al., 1986). Overexpression of the trehalose biosynthetic pathway encoded by *otsA* and *otsB* has previously been shown to improve growth in the presence of osmotic stress (due to salts and sugars) or exogenous ethanol in *E. coli* without ethanol production (Purvis et al., 2005); however, while the minimum inhibitory concentration (MIC) for glucose, xylose, and sodium was increased, the MIC for ethanol was not increased by *otsA-otsB* overexpression and/or betaine supplementation. Ethanologenic *E. coli* overexpressing a heterologous global regulator of radiation-resistance in *Deinococcus radiodurans* has improved osmotic tolerance and ethanol production, and this strain was found to have increased copies of the *otsA* and *otsB* transcripts and increased intracellular trehalose content (Ma et al., 2011). Although reasonable given the effects of trehalose discussed above, to our knowledge, overexpression of *otsA-otsB* has not previously been shown to improve ethanol production in *E. coli*.

4.4.3 Testing ethanol tolerance and ethanol production of selected clones

To test for improved ethanol tolerance and production of the selected clones, we randomly chose 10 colonies from a plated sample from the end of the selection. These 10 clones
(designated A-J) were sequenced, and half of the clones contained genomic fragments in the *uspC-otsA-otsB* genomic region with all of these clones containing the *otsA* and its promoter (Figure 4-4A). The other five clones were found to all contain different genomic regions, and most (4/5) contained small genomic fragments less than 500 bp in length. We also noticed that repetitive extragenic palindromic (REP) elements were present in two clones, and these 30–35 bp elements (~700 annotated in MG1655) have been shown to stabilize mRNA, bind DNA gyrase and polymerase I, and allow for recombination or transposition between REP sites (Gilson et al., 1990; Keseler et al., 2011; Stern et al., 1984).

We next tested each of these ten selected clones for improved growth under ethanol stress and ethanol production in batch fermentations. We chose one colony from a plate of our ethanol adaptations as a control to test in parallel (designated Adapted). Here, ethanol tolerance was quantified as the growth in the presence of 15 g/L or 25 g/L of exogenously added ethanol in otherwise ethanol production conditions. All clones containing part of the *uspC-otsA-otsB* genomic region (B, C, F, I, J) showed improved growth in 15 g/L ethanol (*p*<0.05), up to 32% improvement, and clones C and F showed statistically improved growth of 70–80% in 25 g/L ethanol (*p*<0.1), as compared to the parent (LW06 empty vector) (Figure 4-4B) (absolute values are presented in Figure 4-5). The addition of 15 g/L and 25 g/L of exogenous ethanol corresponds to approximately 80% and 95% reduction in growth (OD$_{600}$ at 24 hr) for the parent, respectively, as compared to no ethanol addition (35% and 60% reduction in specific growth rate on average). The other 5 clones (A, D, E, G, H) and the adapted strain showed no improvement in ethanol tolerance under the same conditions (Figure 4-4B).
A

<table>
<thead>
<tr>
<th>Clone</th>
<th>Start genomic position</th>
<th>Stop genomic position</th>
<th>Insert size (bp)</th>
<th>Genes and extragenic sites</th>
<th>uspC-otsA-otsB region genomic fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4,236,199</td>
<td>4,238,312</td>
<td>2,114</td>
<td>yjbH, yjbT, REP314</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1,977,480</td>
<td>1,979,696</td>
<td>2,217</td>
<td>uspC, otsA, otsB*</td>
<td></td>
</tr>
<tr>
<td>C**</td>
<td>1,977,782</td>
<td>1,981,126</td>
<td>3,345</td>
<td>uspC*, otsA, otsB, araH*</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>3,648,464</td>
<td>3,648,905</td>
<td>442</td>
<td>arsC*, REP262a, REPv262b</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>3,627,172</td>
<td>3,627,403</td>
<td>232</td>
<td>rbbA*</td>
<td></td>
</tr>
<tr>
<td>F**</td>
<td>1,977,782</td>
<td>1,981,126</td>
<td>3,345</td>
<td>uspC*, otsA, otsB, araH*</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>408,470</td>
<td>408,949</td>
<td>480</td>
<td>rdgC*</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>2,906,910</td>
<td>2,907,097</td>
<td>188</td>
<td>pyrG*</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1,977,685</td>
<td>1,979,698</td>
<td>2,014</td>
<td>uspC, otsA, otsB*</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1,977,782</td>
<td>1,980,799</td>
<td>3,018</td>
<td>uspC*, otsA, otsB, araH*</td>
<td></td>
</tr>
</tbody>
</table>

*denotes partial gene
**clones C and F contain identical genomic fragments

B

Graph showing growth in ethanol (OD600 at 24 hr relative to parent) for clones A to J at 15 g/L and 25 g/L EtOH.
Figure 4-4 Genotypes and ethanol tolerance of selected colonies. (A) Ten colonies were picked randomly from the plated selection culture at the end of eighth batch, and their genomic fragments contained on the plasmid were sequenced as indicated here. Five of the ten colonies contained fragments in the uspC-otsA-otsB genomic region, and these fragments are depicted in the right column. Partial genes are denoted (*), and clones C and F contain identical genomic fragments (**). (B) Selected colonies were grown in 15 g/L and 25 g/L of ethanol for 24 hr in 15 ml screw-cap tubes and otherwise identical to the selection environment. Only clones containing the fragments of the uspC-otsA-otsB genomic region show improved ethanol tolerance. Error bars represent the standard deviation from three biological replicates. P-values for comparison to the parent were calculated by a Student’s t-test (* p<0.05 in 15 g/L, ** p<0.1 in 25 g/L).

Figure 4-5 Absolute values of growth for selected clones grown in 15 and 25 g/L ethanol (corresponding relative values reported in Fig. 3B).

Next, we carried out batch fermentations in production medium using these clones to assess any effect on ethanol production (performed in triplicate). For the clones containing fragments of the uspC-otsA-otsB genomic region fragments, clones B and I, which both contain uspC and otsA, had increased growth throughout the fermentation (Figure 4-6A) and increased ethanol titers at 24 hrs (up to about 80 %) along with clones C and F (p<0.05) compared to the parent (Figure 4-6C). Although clones B and I had increased cell densities throughout the fermentation and similarly increased productivities, they did not produce a higher final titer of ethanol. Interestingly, clone J did not show improved growth or ethanol production during the
fermentations despite having improved tolerance to exogenous ethanol under production conditions. The other clones also showed no improvement in growth (Figure 4-6B) or ethanol produced (Figure 4-6D) during the fermentations.

Figure 4-6 Ethanol production of selected clones. Batch fermentations in shake flasks were performed identical to the selection environment except that no exogenous ethanol was added initially. A sample was taken from each fermentation every 24 hr, and the cell density and ethanol concentrations were measured for all samples. Growth during the fermentation relative to the parent strain for (A) the clones containing the uspC-otsA-otsB genomic region and (B) all other selected clones is plotted. Ethanol production during each fermentation relative to the parent strain for (C) the clones containing the uspC-otsA-otsB genomic region and (D) all other selected clones is plotted. The average values for the empty vector strain are in parentheses next to the axis. Error bars represent the standard deviation of three biological replicate fermentations.
We found no improvement in ethanol tolerance or production for the adapted strain unlike previous adaptation studies, and this difference is likely due to the relatively short period of selection (8 days) compared to the typical 50-90 days of adaptation studies. While we cannot explain fully why certain clones present at the end of our selection did not show improved ethanol tolerance or production (A, D, E, G, H), there are a few potential explanations worth elaborating. Since these clones are present after many generations of selection (33 generations) and these genomic regions were not found at high initial frequencies in the libraries, (all were below average frequency on the MAS5.0 Affymetrix CHP file and determined by SCALEs), we suspect that these clones were indeed selected for but that the conditions under which they were selected for have changed in some aspect. One possibility is that the genetic makeup of these clones changed at some point during the selection such that they were enriched early in the selection and then changed genetically at some later time during the selection. These genetic changes could result from homologous recombination between plasmids and the chromosome, particularly in the case of the clones with REP elements being present. Our LW06 ethanol platform strain is recA+ which may play a role in this genetic transfer. Another possibility is that the initial clone contained multiple unique plasmids, and in the case of a double transformation if the larger plasmid conferred improved tolerance, the smaller plasmid with the identical origin of replication could be replicated more quickly and eventually displace the larger plasmid over many divisions. A third possibility is that the presence of these clones resulted from cross-feeding with other clones present in the selection or relied upon using cell components from lysed cells that would no longer be present during the confirmation tests, but we believe this to be less probable given that most of these clones contain only partial gene fragments that do not encode for full gene products and therefore are probably not involved in metabolism.
Here we have shown that we can indeed select for clones with improved ethanol tolerance and ethanol production. Additionally, we found that we identified clones with improved growth under conditions identical to the selection (15 g/L ethanol), and most (4 out of 5) of these clones had improved productivities of ethanol over the first 24 hours of fermentation. We next performed sub-cloning analysis of the uspC-otsAB region to identify individual and combinatorial contributions of each of these genes on growth under high ethanol stress and ethanol production.

4.4.4 Combinatorial analysis of ethanol tolerance and ethanol production genes

Since the clones which contained uspC and otsA showed the largest improvement in ethanol production, we first sought to determine if UspC (universal stress protein) specifically contributed to ethanol tolerance, since unlike OtsAB, UspC has not previously been associated with resistance to ethanol or osmotic stress. To do so, we constructed six clones to combinatorially test overexpression of the genes in the uspC-otsA-otsB genomic region (genomic fragments containing entire gene transcripts were cloned). We tested the growth of these clones in 25 g/L to investigate their ethanol tolerance (Figure 4-7A). Individually, only otsA conferred ethanol tolerance alone, and all clones containing otsA also had statistically improved growth ($p<0.05$), which suggests that otsA is required for the ethanol tolerance phenotypes identified here. Although otsB alone did not confer ethanol tolerance, the addition of otsB in the otsA-otsB clone further improved growth under ethanol stress as compared to otsA alone ($p=0.025$) suggesting that otsA might be the rate-limiting enzyme in the trehalose biosynthesis pathway, as has been previously suggested (Padilla et al., 2004). In this case, otsB overexpression would only improve ethanol tolerance when otsA is overexpressed and no longer the rate-limiting step.
For the *otsA-otsB* clone here, overexpression of *otsB* should also further increase transcription of *otsA* as compared to the *otsA* clone since the *otsB* promoter initiates transcription of the *otsBA* operon (see Figure 4-4A schematic). Therefore, either increased expression of *otsA* or the addition of *otsB* overexpression could be responsible for the improved ethanol tolerance in the *otsA-otsB* clone as compared to the *otsA* clone. In the presence of high ethanol stress, the *otsA-otsB* clone had the most growth of these constructed clones, which may explain why the majority of the selected cloned containing the *uspC-otsA-otsB* genomic region contained genomic fragments encoding the full *otsA* and *otsB* transcripts (3 out of 5 clones: C, F, J).

We next tested these constructed clones for ethanol production in batch (shake flask) fermentations, and similar to the ethanol tolerance findings, all clones containing *otsA* had improved growth and productivity of ethanol over the first 24 hours of fermentation, up to 50% improvement (Figure 4-7B, Figure 4-7C). Additionally, as in the selected clones, the constructed clones that have improved rates of ethanol production (*otsA, uspC-otsA, otsA-otsB, uspC-otsA-otsB*) do not have improved final titers of ethanol. Interestingly, although *uspC* in combination with *otsA* did not improve ethanol tolerance, during fermentation the addition of *uspC* along with *otsA* improves the concentration of cells throughout the fermentation as compared to *otsA* alone (*p<0.15*), yet this increase in cell mass for the *uspC-otsA* clone did not correspond with an increase in ethanol produced after the first 24 hr.
Figure 4-7 Ethanol tolerance and ethanol production of constructed clones to combinatorially test genes in the uspC-otsA-otsB genomic region. (A) Growth (OD600) in the presence of 25 g/L of ethanol (and otherwise production conditions) after 24 hr was measured for biological triplicates and compared to the empty vector (* p<0.05). Batch fermentations in shake flasks were performed and the (B) cell growth (OD600) and (C) ethanol produced relative to the empty vector parent strain are plotted, and the average values for the empty vector strain are in parentheses next to the axis. Error bars represent the standard deviation of three biological replicate experiments.
The *uspC-otsA* clone also had the greatest average growth of all of the clones during fermentation, including more than *otsA-otsB* which showed the most ethanol tolerance, suggesting that *uspC* is synergistic with *otsA* for osmotic stress tolerance (fermentations have undetectable ethanol initially present <0.1 g/L) but not for higher level ethanol tolerance (25 g/L). UspC would be expected to have a role in osmotic stress tolerance under this high K⁺ environment since it is involved in the regulation of intracellular K⁺ during osmotic stress (Heermann et al., 2009). Using our genome-wide approach, we identified the *otsA-otsB* (3/5 clones) and *uspC-otsA* (2/5 clones) genotypes conferring improved growth under the selective condition modeled after ethanol production conditions, and here we see that these genotypes are tailored to ethanol tolerance or osmotic tolerance with intracellular ethanol production.

### 4.5 Conclusions

Here we have demonstrated a genome-wide approach using a production platform strain to perform complete genome mapping for genotypes related to improved production of ethanol. Using this approach we selected for clones conferring both improved ethanol tolerance and production of ethanol. The *uspC-otsA-otsB* genomic region was most highly enriched in the selection and the selected clones containing fragments of this genomic region conferred these improved traits. We performed further sub-cloning analysis to test the individual and combinatorial effects of these genes in relation to ethanol tolerance and production. We identified that *otsA* predominantly confers the improved ethanol tolerance and productivity phenotypes, the addition of neighboring *otsB* further improves ethanol tolerance, and the neighboring *uspC* further improves ethanol productivity. Interestingly, both of these genotypes (*otsA-otsB, uspC-otsA*) were clones directly identified in our selection. Future work could utilize the genetic targets identified here along with the additional genome engineering approaches
(Lynch et al., 2007; Wang et al., 2009; Warner et al., 2010b) to engineer further improvements to the production of ethanol (or possibly other alcohols), such as selecting for improved MIC or combinatorial optimization of the expression levels for the genetic targets identified here. The approach that we have demonstrated here could also be more generally applied to other production systems in which production of an inhibitory product is growth-related in order to select for improved production-related traits.

Acknowledgments

We thank Robert Dong for assistance with library construction. LBA Woodruff was supported by a NSF Graduate Research Supplement (CBET0449183).

4.6 References


Chapter 5 Future Work: Engineering trait-specific regulators using libraries of parts to improve *E. coli* ethanol production

5.1 Introduction

Microbial cells can be genetically reprogrammed to produce a range of products, including fuels and commodity chemicals. Yet, to achieve commercially relevant performance these biological factories often require additional genetic engineering of complex and poorly understood phenotypes (Dunlop, 2011; Nicolaou et al., 2010). Stress tolerance is a key issue in this regard, where production strains must perform in the presence of toxic products, intermediates, and feedstocks as well as under non-ideal osmolarity, temperature, or pH. Stress tolerance has traditionally been improved by long-term adaptation, which relies upon random mutagenesis across the entire genome. A number of approaches have been developed to accelerate strain engineering by enabling the mapping of genotype-phenotype relationships in parallel for the entire genome and thus, directing the genetic search space to the most relevant collection of mutations (Borden and Papoutsakis, 2007; Giaever et al., 2002; Gill et al., 2002; Goodarzi et al., 2010; Lynch et al., 2007; Smith et al., 1995; Warner et al., 2010a).

A variety of strategies and approaches to engineer complex traits have been previously reported and reviewed (Patnaik, 2008; Santos and Stephanopoulos, 2008; Woodruff and Gill, 2011). Since the expression of complex traits is sensitive to a variety of genetic and environmental factors, engineering of such traits is fundamentally bounded by the realities of combinatorial mathematics and limitations in laboratory throughputs. One creative strategy for addressing this issue has involved the engineering of regulatory proteins, which elicit cellular-wide expression changes by regulating transcription or translation of multiple genetic targets. In
fact, engineered transcription regulators, such as artificial transcription factors containing zinc fingers (Lee et al., 2008; Park et al., 2003) or randomly mutated global regulators by global transcription machinery engineering (gTME) (Alper et al., 2006; Alper and Stephanopoulos, 2007), have been used to improve a number of industrially-relevant traits in eukaryotic and bacterial systems.

The gTME approach has proven to be a versatile approach that has been used to engineer various phenotypes in *E. coli* by screening mutant libraries of *rpoD* and *rpoA* (Alper and Stephanopoulos, 2007; Klein-Marcuschamer et al., 2009; Klein-Marcuschamer and Stephanopoulos, 2008; Yu et al., 2008). Nevertheless, selecting a global regulatory protein to engineer a specific phenotype has often required screening libraries of multiple global regulators, and the most improved phenotypes have typically resulted from only a few mutations (≤ 5) or truncated regulatory proteins (Klein-Marcuschamer et al., 2009; Lanza and Alper, 2012). Additionally, regulator libraries using the gTME approach have been constructed by random mutagenesis which can search only a small fraction of the genotypic variants (~10^6 clones out of the 10^{390} amino acid variants for a gene of 300aa), yet only <1 % of point mutations within a protein are typically beneficial (Romero and Arnold, 2009). With recent advances in DNA synthesis, genetic diversity can now be designed to target specific functional modifications, such as within specific residues of an enzyme in protein engineering or within libraries of parts in synthetic biology applications.

Here we demonstrate a new regulator engineering approach which is possible due to these recent advances and which addresses the above limitations with the following modifications: (i) a trait-specific regulatory protein is identified by mapping quantitative, genome-wide genotype-to-phenotype relationships for the trait of interest onto the regulatory network; (ii) genetic
diversity within the regulator is targeted to parts of the regulatory protein based on the desired effect on cellular expression changes; (iii) mutations within these parts are designed to alter specific functions while allowing for all genetic variants to be tested in the limited search space \(10^6\); (iv) all introduced mutations can be tracked during growth in the selective environment using high-throughput sequencing; and (v) combinatorial optimization of the mutated regulator to further improve the trait during additional rounds is informed by the quantitative fitness for all mutations in the library (Figure 5-1). While this is a novel approach to engineer regulatory proteins for improved complex phenotypes, notably, most elements of this approach are inspired by and have been used to engineer improved enzymes by directing the search space to functional mutations (Bloom et al., 2006; Chen et al., 2012; Feng et al., 2012; Fox et al., 2007; Herman and Tawfik, 2007; Lutz, 2010; Reetz and Carballeira, 2007; Reetz et al., 2008; Romero and Arnold, 2009; Voigt et al., 2002).

In this work, we are applying this approach to engineer trait-specific regulators by libraries of designed parts to improve the industrially-relevant system of ethanol biofuel production in \textit{E. coli}. In a previous study, we used our LW06 ethanol production strain (ATCC BAA-2466) to perform a genome-wide study to quantitatively map genotypes related to \textit{E. coli} ethanol production (Woodruff et al, submitted). Here, utilizing this data, we present approaches to identify transcriptional regulators relevant to this system, and using custom oligonucleotides to introduce targeted genetic diversity, we test multiple library designs through competitive selective under ethanol production conditions.
Figure 5-1 The trait-specific regulon engineering by assemblies of parts approach. (i) Data from a quantitative, genome-wide search for the desired trait, in this case a genome-wide selection for genotypes related to ethanol production in E. coli (Woodruff et al, submitted), (ii) is analyzed for trait-specific regulatory targets, for example by mapping the fitness onto the regulatory network. (iii) The identified trait-specific regulator is engineered by designing libraries which introduce genetic diversity into parts of the regulator which focus the genetic search space to mutations likely to alter cellular expression related to the trait of interest, for example by targeting the ribosomal binding site or the annotated DNA binding site with complete library coverage. (iv) Libraries of the engineered regulator are constructed by assembling the designed parts, subjected to competitive growth in a selective environment (or screened for the desired trait), and the fitness of the introduced mutations are quantified by high-throughput sequencing. The regulator can be recursively engineered for further strain improvements by designing successive libraries of parts based on the fitness landscape of the tested parts libraries.

5.2 Materials and Methods

5.2.1 Strains, media, and reagents

The LW06 E. coli ethanol production strain (BW25113 ΔldhA ΔackA ΔfrdABCD ΔadhE attTn7::P_{LlacO-1} pdc_{Zm} adhB_{Zm} AmpR) was used for all experiments (Woodruff et al, submitted). This strain is available from the ATCC (# BAA-2466). AMX minimal media used for the selections, fermentations, and growth testing, where AMX is a modified version of the AM1 media previously optimized for ethanol and lactate production (Martinez et al., 2007) which does not contain any betaine supplementation and additionally contains 0.1 M MOPS buffer (1M stock solution buffered to pH 7.2 with 10 M KOH) and 0.1 M monobasic potassium phosphate (1M stock solution buffered to pH 7.4 with 10 M KOH), otherwise identical recipe to AM1.
Buffers and concentrations were experimentally determined to maximize batch ethanol production of LW06 at 100 hr with a final pH ≥ 6. AMX media was freshly prepared prior to use. Carbenicillin (100 μg/ml) was added to all LW06 cultures, and kanamycin (30 μg/ml) was added to all cultures containing the pSMART-LCKan vector. Molecular biology grade ethanol was used for ethanol dosing (Sigma-Aldrich E7148). Sequencing was performed by Operon. Oligos and primers were synthesized by IDT.

5.2.2 Clone construction

Primers were designed to amplify the wild-type regulators of interest from *E. coli* MG1655 genomic DNA (ATCC 47076): *ycjW* _1380961_f (5′–tggcattcaggcaaggtg−3′), *ycjW* _1382100_r (5′–tgcgtaggtatgatgtgggg−3′), *dctRslp* _3651910_f (5′–tcacctcagaatcagatgaaaacta−3′), *dctR*_ _3653264_r (5′–gtagccagactcacegtgta−3′), *rpoE*_ _2707442_f (5′–ccagtatcccgctatcgtca−3′), *rpoE*_ _2708272_r (5′–cgttgggttactcttcagggca−3′), *rpoH*_ _3597951_f (5′–attacgcttcaatggcagcac−3′), *rpoH*_ _3599075_r (5′–tttaacgccactttacgc−3′). These primers were phosphorylated and used to amplify the desired genomic region using Phusion high-fidelity DNA polymerase (NEB) according to the manufacturer’s recommendations. PCR products were purified from a gel using a QIAquick Gel Extraction kit (Qiagen) and ligated to the pSMART-LCKan backbone using the CloneSMART kit (Lucigen) by incubation overnight at 4 °C and 2 hr at 25 °C. Following heat inactivation, each ligation reaction was transformed into LW06 made electrocomponent by ultrapure water washing. Transformations were plated on LB with kanamycin to select for clones. Plasmid inserts were verified by gel electrophoresis following digestion with SwaI to check for chimeras and plasmid insert sequences were verified by sequencing with primers SL1 (5′–cagtccagttacgctgagtc−3′) and SR2 (5′–gtgtagtatgattaatggct−3′).

5.2.3 Library construction
The libraries of the regulators were constructed using overlap-extension PCR to assemble the parts of the regulator. The overlapping regions for each part were designed to have a melting temperature of about 60 °C under PCR conditions. Wild-type parts and linker regions were amplified using purified plasmid DNA containing the constructed wild-type regulon as the template and purified by gel extraction using a QIAquick Gel Extraction kit (Qiagen) prior to PCR assembly. Library oligos containing the desired mixed bases in the RBS (~60 bp oligos) and DNA binding site (~110 bp ultramers) were ordered as PAGE-purified oligos or ultramers from IDT. Library oligos were amplified and purified using a QIAquick PCR Purification Kit (Qiagen). Amplified DNA parts were assembled piecewise using overlap-extension PCR with the appropriate flanking primers and 1 pmol of each of the two DNA parts. DNA concentrations were quantified by their A_{260}. All PCRs were performed using Phusion polymerase (NEB) according to the manufacturer’s recommended protocol with Phusion HF reaction buffer, 3 % DMSO, a 30 sec 55 °C annealing step, and 30 cycles of amplification.

5.3 Results and Discussion

5.3.1 Identification of trait-specific regulators

The choice of the regulatory target is a key factor in determining the cellular expression states accessible through engineering. Further, there is a perceivable tradeoff between the breadth of genes regulated by a given target and the capability of combinatorial optimizing or fine-tuning expression of only the subset of gene products relevant to the desired phenotype rather than the larger set of regulated genes by further mutation of the regulator. This could explain the typical small number of mutations and truncated proteins in the improved mutated global transcription factors when targeting \( rpoD \) and \( rpoA \), which are necessary for housekeeping expression of most or all genes, respectively (Lanza and Alper, 2012). Additionally,
previous studies using the gTME approach have found that not all randomly mutagenized global regulator libraries tested have resulted in eliciting the desired phenotype (Klein-Marcuschamer et al., 2009; Yu et al., 2008). Here, we applied a genome-wide approach to map overexpression of all genes, including all transcription elements, onto the cellular regulatory network in order to identify four trait-specific regulators related to the highest fitness genes in our E. coli ethanol production system.

In a previous study, we used the multiscalar analysis of library enrichments (SCALEs) genome-wide approach to identify genomic regions conferring improved ethanol production from high resolution genomic libraries (Woodruff et al, submitted). We analyzed the previously calculated gene fitness values for genes quantified by selection and involved in transcription (annotated to COG K), which resulted in 214 genes (Figure 5-2A). Two predicted DNA-binding transcriptional regulators, dctR and ycjW, were the most enriched transcription regulators midway through the selection (blue) and at the end of the selection (red), and they were also among the highest fitness genes identified out of all genes (gene rank ≤10 at midway, rank ≤27 at end). DctR is involved in resistance to formic, lactic, and succinic acid in spent LB media at low pH (2.5) (Mates et al., 2007) and overexpression causes filamentous morphology associated with biofilm formation (Tenorio et al., 2003). YcjW has been shown to mitigate the lethal effects of stress of the quinolone nalidixic acid and is believed to be involved in the bacterial stringent response (Han et al., 2010). The genes regulated by DctR and YcjW are currently unknown.
Figure 5-2 Identification of regulatory targets related to ethanol production in *E. coli*. (A) The gene fitness values (peak height) from our previous genome-wide selection for ethanol production (Woodruff et al, submitted) for all genes related to transcription (COG K) are plotted at their corresponding genomic location for the fourth batch (blue inner circle) and final eighth batch (red outer circle) of selection. (B) The 500 genes with the greatest fitness at the end of the previous selection were mapped onto the regulatory network according to their regulators in order to identify global regulators that regulate expression of these genes related to ethanol production. The regulatory networks of the global regulators (C) *rpoH* (σ^{32}), (D) *rpoE* (σ^{24}), (E) *rpoS* (σ^{38}), and (F) *rpoD* (σ^{70}) are shown for reference with the number of genes regulated for each in parenthesis. Regulatory network diagrams were constructed using the EcoCyc regulatory overview (Keseler et al., 2011).
In addition to identifying transcription regulatory targets based on their direct enrichment during selection, we also sought to identify global regulatory targets which regulate the most relevant genes for the desired phenotype. Therefore, we mapped the highest fitness genes (500) identified in our previous selection onto the regulatory network in order to identify global regulators involved in expression of genes relevant to improved *E. coli* ethanol production (Figure 5-2B). These ethanol production related genes are distributed throughout the regulome; however, genes are clustered in the regulatory networks of the sigma factors *rpoH* (Figure 5-2C), *rpoE* (Figure 5-2D), *rpoS* (Figure 5-2E), and *rpoD* (Figure 5-2F). While engineering any of these regulatory targets could result in improved ethanol production characteristics, here we chose to engineer the *rpoH* and *rpoE* sigma factors over *rpoD* and *rpoS*. The Eσ70 holoenzyme of *rpoD* has a vast regulatory network as the housekeeping sigma factor which regulates transcription of approximately 50% of all genes and could present challenges for combinatorially optimizing expression of a phenotypically relevant subset of genes by altered promoter binding specificities (Figure 5-2F). The *rpoS* stress response sigma factor is expressed at very low levels during exponential growth (Battesti et al., 2011), which we have during the growth selections and has the weakest binding affinity for the RNA polymerase core enzyme (Maeda et al., 2000) which in combination could limit the phenotypic diversity in our system. Both *rpoH* and *rpoE* are sigma factors involved in the response to heat shock (Erickson et al., 1987; Grossman et al., 1984). In yeast, the stress responses to sublethal ethanol and heat shock are highly similar (Piper, 1995). In addition to thermal stress, *rpoH* is also induced by ethanol stress (VanBogelen et al., 1987) and rpoE to structural changes of lipopolysaccharides in the outer membrane (Tam and Missiakas, 2005). Moreover, *rpoH* and *rpoE* are also induced by hyperosmotic stress (Bianchi and Baneyx, 1999), and overexpression of the trehalase biosynthesis pathway, which is
known to confer osmotic tolerance, was previously found to confer improved ethanol tolerance and production in this system (Woodruff et al, submitted).

5.3.2 Regulon library design and construction

Although this regulatory engineering approach could be used to target any region of the regulon, here we chose to focus the search space to mutations within the ribosomal binding site (RBS) and DNA binding site parts in order to tune the expression of the regulatory protein and its promoter binding specificity. In this work, we are testing our hypothesis that mutations within the DNA binding site which alter promoter binding affinities will be more prevalent for engineered clones of the sigma factor libraries, which were identified as targets based on enrichment of their regulatory networks as opposed to direct enrichment in the selection. Additionally, we hypothesize that tuning expression levels through mutated RBSs to be found for all regulators. Bacterial sigma factors are transcription initiation factors that are primarily responsible for promoter recognition, and the resulting transcription pattern of the RNA polymerase transcription machinery is determined by which sigma factor is contained in the holoenzyme (Gruber and Gross, 2003; Ishihama, 2000). The 7 sigma factors of *E. coli* compete to bind to the RNA polymerase core enzyme, and while they differ about 16-fold in binding affinity ($\sigma^D > \sigma^N > \sigma^F > \sigma^H > \sigma^E > \sigma^S$), changing the ratio of the sigma factors present has been shown to alter the distribution of the holoenzyme species (De Vos et al., 2011; Maeda et al., 2000). All four of the identified regulatory targets (*ycjW*, *dctR*, *rpoE*, and *rpoH*) contain annotated helix-turn-helix (HTH) DNA binding motifs (Keseler et al., 2011). HTH motifs are prevalent in bacterial one-component transcription factors (84%) (Ulrich et al., 2005) and are comprised of two $\alpha$-helices joined by a sharp turn linker, the second helix of which is the DNA “recognition helix” that inserts into the major groove of the DNA allowing complex hydrogen
bonding and van der Waals interactions between the protein side chains and DNA bases (Brennan and Matthews, 1989). Studies on mutant rpoD sigma factors have shown that mutations contained within the HTH motif, which is responsible for recognition of the –35 promoter region, alter the promoter binding specificity of the resulting RNA polymerase holoenzyme (Gardella et al., 1989; Kim et al., 1995; Siegele et al., 1989).

For these reasons, we are targeting genetic diversity to the RBS and HTH DNA binding region parts of the four identified regulators, yet we further designed the introduced mutations and libraries to ensure complete coverage of the genetic variants of these parts within our libraries (10^6 each). The degenerate RBS is designed to resemble the Shine-Dalgarno conical sequence identical to the design previously implemented (Wang et al., 2009) and has about 2×10^4 possible variants (Figure 5-3A). Ten DNA binding cassettes were designed for each regulator to test site-saturated mutagenesis of all 20 HTH residues with 2 NNK mutated codons maximally distanced nine positions apart in each oligo for a total of 1×10^4 variants (Figure 5-3A). By using multiple DNA binding site cassettes, we can select from all variants, as opposed to 10^{30} variants if fully degenerate which drastically exceeds library size limits, and have a higher a fraction of functional cassettes which retain the HTH motif. We chose NNK codon mutagenesis which reduces the fraction of stop codons by 50% over complete degeneracy yet encodes for all 20 amino acids, though not in equal proportions (9%: L,R,S; 6%: A,G,P,T,V; 3%: stop, 12 remaining amino acids). From these parts, we are constructing 3 libraries for each regulator (Figure 5-3B) in order to test all RBS and DNA binding site variants in the libraries with only one set of library oligos, and in the third library we are testing only a small fraction of the combinatorial variants for these two parts to investigate the interactions between these part
libraries. Probabilities of library completeness were calculated using GLUE and GLUE-IT (Patrick and Firth, 2005).

Figure 5-3 Design of the libraries of regulatory parts. (A) Genetic diversity was introduced into the regulator’s ribosomal binding site (RBS) and DNA binding site using custom oligos constructed with the desired mixed bases. The degenerate RBS design is identical to that used by MAGE (Wang et al., 2009), which introduces the degenerate RBS sequence (D=A,G,T; R=A,G) at the -4 to -14 positions upstream of the start codon. The 20 amino acid helix-turn-helix motif comprising the annotated DNA binding site of each targeted regulator was mutagenized by NNK (N=A,T,G,C; K=G,T) site saturated mutagenesis with each of the 10 custom oligos each containing two mutagenized codons the maximally distanced 9 residues apart. (B) Three libraries for each regulator were constructed to completely search the genotypic space of the RBS and DNA binding site parts as well as investigate some of the interactions between these parts with a combinatorial library that searches a small fraction of the combinatorial variants.
5.3.3 Future work

We have successfully tested the assembly of the regulator libraries using the designed PCR amplified parts and the overlap-extension PCR assembly method. Additionally, we have cloned the four parent regulator clones and verified them by sequencing. Future work scheduled for the next few months will entail finishing the regulon assemblies by PCR, the library constructions using identical cloning methods as were used for the genomic library construction and then performing the selections under ethanol production conditions nearly identical to our previous ethanol production selections with the exception that here we will mix equal proportions of the 3 libraries for each regulator (Figure 5-3B) in place of the genomic libraries. Each of the 4 regulatory targets (ycjW, dctR, rpoE, and rpoH) will be performed as separate serial batch selections. The region containing the RBS to the DNA binding site will be PCR amplified from plasmid DNA extracted from the initial mixed libraries and a couple selection time points for each regulator. These PCR product samples will then be tagged and sequenced at the RBS and DNA binding site locations by paired-end 2 × 100 bp high-throughput using an Illumina HiSeq 2000 sequencer at the BioFrontiers Next-Generation Genomics Facility. Additionally, we will assay 5 selected clones from each selection for ethanol tolerance and ethanol production as performed previously. The results from these experiments will be used to determine our combinatorial optimization strategy. We plan to test further combinatorial optimization for one regulator.
5.4 References


Chapter 6 Conclusions and Future Directions

Engineering improved tolerance to inhibitory products is of broad interest to microbial bioprocesses for a wide range of products (Dunlop, 2011; Nicolaou et al., 2010; Patnaik, 2008). In this collection of research, we have investigated approaches for this purpose using the relatively well-studied system of ethanol tolerance and ethanol production in E. coli as our model system. Nevertheless, using this model system, we have developed genomic approaches to engineer improved ethanol production which could be applied to other production systems and uncovered findings, such as the limitations of commonly used approaches to engineer product tolerance, which are broadly applicable to engineering phenotypes of improved tolerance.

In Chapter 1 (Woodruff and Gill, 2011), we discussed the challenges of directing genetic search space within the limitation of library size to generate meaningful genotype-to-phenotype relationships. In the context of an overall genomic engineering algorithm (genome searching, relevance ranking, and then combinatorially optimizing), we presented genomic engineering approaches that have been recently developed and can be used in concert for genomic engineering of improved traits. In Chapter 2, we carried out the first step of this algorithm, searching the genome for genetic targets, to quantitatively map the entire genome for genetic targets conferring ethanol tolerance in wild-type E. coli. Ethanol tolerance has been widely accepted to be a complex (multigenic) phenotype (Alper and Stephanopoulos, 2007; Brynildsen and Liao, 2009; Gonzalez et al., 2003; Goodarzi et al., 2010), yet in this work, we demonstrated that by using a different approach we could identify individual genes, moreover all novel genetic targets, which confer improved ethanol tolerance when overexpressed.
In Chapter 3, we sought to investigate the relationship between ethanol tolerance and ethanol production using the ethanol tolerance genes identified in Chapter 2. These experiments necessitated an ethanol production \textit{E. coli} strain in place of the wild-type \textit{E. coli} strain used in Chapter 2. Therefore, we constructed an ethanol production platform strain with fully-characterized and chromosomally-integrated genetic modifications for ethanol production since a strain with these characteristics was not available, yet our design was partially based on the previously constructed \textit{E. coli} ethanol production strains (Ohta et al., 1991; Yomano et al., 1998; Yomano et al., 2008). We tested the effect of our best ethanol tolerance genes identified in Chapter 2 on ethanol production and tolerance under various conditions to uncover that neither gene improved ethanol production. Additionally, the ethanol tolerance of these genetically distinct clones was shown to be dependent on the host genotype, metabolism \textit{(i.e., induction of ethanol production)}, and media of the system, which can help explain the previously contradictory studies on the relationship between tolerance and production (Agudo, 1985; Alper et al., 2006; Atsumi et al., 2010; Baer et al., 1987; Dunlop, 2011; Dunlop et al., 2011; Yomano et al., 1998) and emphasizes the importance of engineering improved tolerance within a production system. Since production platforms have limited availability, we presented our LW06 ethanol platform, which we made openly available from the ATCC, in the framework of a suggested metabolic engineering “commons” collection of platform production strains that would enable the greater metabolic engineering community to carry out their engineering efforts and studies within the relevant production system.

Based on finding in Chapter 3 that ethanol tolerance in a wild-type host and standard laboratory conditions does not directly transfer into improved ethanol production, in Chapter 4 we redesigned a genome-wide approach to search the entire genome for genetic targets to
improve ethanol production in our system. Our approach allows for a systematic search of the entire genome, whereas as previous approaches to engineer improved production have relied on random mutagenesis targeted either to the entire genome or the global transcription machinery (Alper et al., 2006; Atsumi et al., 2010; Chen et al., 2011b; Connor et al., 2010; Ma et al., 2011; Smith and Liao, 2011). In this study we selected for clones with both improved ethanol tolerance and improved rates of ethanol production. Through further subcloning we combinatorially tested the contribution of the highest fitness genes and their combinations to improvements in ethanol tolerance and ethanol production. Interestingly, the gene combinations resulting in the largest improvements under each of these conditions were different yet both were the highest frequency genotypes among the selected clones.

We then used the data from our genome-wide search for ethanol production-related genes in Chapter 4 to rank the relevance of regulatory genetic targets and identify trait-specific regulatory targets for a novel regulator engineering approach which we present in Chapter 5. Using our ethanol production platform and selection system presented in Chapters 3 and 4, we will investigate this regulatory engineering approach to engineer and combinatorially optimize complex phenotypes eliciting improved ethanol production. Our approach to engineering trait-specific regulators using libraries of parts is inspired by the gTME approach (Alper et al., 2006; Alper and Stephanopoulos, 2007) as well as advancements in protein engineering and synthetic biology, yet the combination of these elements in the method we present should yield a marked improvement over global random mutagenesis.

Many opportunities for future work that could build on the results of this work exist, some of which I highlight here. In Chapter 2 we identified nine novel genetic targets which individually confer improved ethanol tolerance, and these targets could be tested combinatorially
to further investigate the role of epistasis in engineering combinatorially optimized multigenic traits (Sandoval et al., 2012). In Chapter 3, we investigated the transferability of tolerance phenotypes and the relationship of tolerance and production with a few genetic targets, yet a much greater depth of study with a greater diversity of genetic targets and different (non-ethanol) production systems will be required to fully understand these relationships. We also presented a model strain for a simple production chassis with many desirable characteristics for genomic engineering of desired traits in Chapter 3. Nevertheless, there are many further improvements which could be incorporated into these production platforms, such as optimization of the production pathway enzymes to reduce pathway bottle necks and toxic intermediate accumulation (Dueber et al., 2009; Ro et al., 2006; Wang et al., 2009) and removal of inherent regulatory mechanisms through refactoring of the production pathway (Temme et al., 2012). In Chapter 4 we presented a directed genome-wide approach which selected for clones containing known genetic modifications with improved ethanol tolerance and production, and this approach utilizing genomic libraries could be applied similarly to improve the production of other inhibitory products that are produced from metabolites involved in cell growth, which includes a wide variety of production systems including other desirable biofuels and commodity chemicals (Atsumi et al., 2008; Lee, 1996; Lu et al., 2008). We also mentioned in Chapter 4 that selection conditions could be designed to search for improved ethanol titers instead of or along with improved ethanol productivity (rate of production). Previous work from our group investigating selection design (Gall et al., 2008b; Warnecke et al., 2008) would suggest that this might be achieved by selections performed at higher concentrations and possibly in combination with successive batches of selection with a decreasing ethanol gradient if growth and not purely survival is also desired. The approach to engineer trait-specific regulators using libraries of parts
which we present in Chapter 5 has the potential to be a versatile approach to engineer complex phenotypes and will be an additional tool in the metabolic engineering toolbox.

6.1 References


Bibliography


