Degenerate oligonucleotide primed amplification of genomic DNA for combinatorial screening libraries and strain enrichment

Benjamin Gordon Freedman

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Biological Systems Engineering

Ryan S. Senger (Chair)

Justin R. Barone

Chenming (Mike) Zhang

Guillaume Pilot

November 13, 2014

Blacksburg, VA

Keywords: Metabolic Engineering, Genomic DNA Library, Degenerate Oligonucleotide Primed PCR, Whole Genome Amplification, Raman Spectroscopy, *Clostrdium cellulolyticum*

Degenerate oligonucleotide primed amplification of genomic DNA for combinatorial screening libraries and strain enrichment

Benjamin Gordon Freedman

ABSTRACT

Combinatorial approaches in metabolic engineering can make use of randomized mutations and/or overexpression of randomized DNA fragments. When DNA fragments are obtained from a common genome or metagenome and packaged into the same expression vector, this is referred to as a DNA library. Generating quality DNA libraries that incorporate broad genetic diversity is challenging, despite the availability of published protocols. In response, a novel, efficient, and reproducible technique for creating DNA libraries was created in this research based on whole genome amplification using degenerate oligonucleotide primed PCR (DOP-PCR). The approach can produce DNA libraries from nanograms of a template genome or the metagenome of multiple microbial populations. The DOP-PCR primers contain random bases, and thermodynamics of hairpin formation was used to design primers capable of binding randomly to template DNA for amplification with minimal bias. Next-generation high-throughput sequencing was used to determine the design is capable of amplifying up to 98% of template genomic DNA and consistently out-performed other DOP-PCR primers.

Application of these new DOP-PCR amplified DNA libraries was demonstrated in multiple strain enrichments to isolate genetic library fragments capable of (i) increasing tolerance of *E. coli* ER2256 to toxic levels of 1-butanol by doubling the growth rate of the culture, (ii) redirecting metabolism to ethanol and pyruvate production (over 250% increase in yield) in *Clostridium cellulolyticum* when consuming cellobiose, and (iii) enhancing L-arginine production when used in conjunction with a new synthetic gene circuit.

Acknowledgements

This project owes its success to many people. First and foremost, I must thank Ryan Senger for giving me the opportunity to come to Virginia Tech and mentoring me along from a fresh graduate to where I am today. Most importantly, his perseverance towards the research goals and ability to find the silver lining within what I had considered "failed" experiments had been the saving grace, which kept me, returning to the lab day after day.

I also need to thank the teachers I have had the pleasure of working with over my career. First off, my teachers and dissertation committee members Justin Barone, Mike Zhang, and Guillaume Pilot who have held me to a higher standard than I held myself. I also acknowledge Hadi Nazem Bokaee, Theresah Zu, Ahmad Athamneh, Imen Tanniche, and the undergraduate researchers, summer students, and collaborators who have come and gone. I have had a pleasure of working aside you all these past 5 years.

I owe my parents, Dale and Jon, a great deal, along with my grandpa Eli and grandma Aldene for their love and support, phone calls, care packages and visits to their only son and grandson to reach below the Mason Dixon.

I thank my good friends from the Outdoors Club at Virginia Tech who kept my weekends interesting and showed me the confidence to reach outside my comfort zone. I owe you all for my tool collection, my chainsaw license, and that one run-in with the Law. Especially my good friends NL, CS, DC, DA, B&HC, FG, NJ, R&NG, EP, JB, RV, BO, DW, LH, PM, BL,CB, JT, EK, TC, DR, my *in loco parentis* R&LB and SP&RL, the Blacksburg 9 softball team, the Steuer Clan, and the hundreds of others over the years who I remember to this day.

Finally there is Katie, who stressed as much through this degree as I did but kept pushing forward. From making coffee in the mornings to packing lunches to telling me I can't wear black

shoes with a navy suit, you've been my career coach, my moral support, my brunch companion, co-pilot, and love of my life.

Thank you,

Benjamin G. Freedman

Table of contents

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Motivation

Humans depend on more materials, chemicals, and energy now than ever before. These chemicals are used to make up practically everything we associate with from the clothes we wear, the cars we drive, the food we eat, the medicines doctors prescribe when we are sick, all the way to the headphones that play our music. Petroleum is responsible for much of our luxury, as the chemical feedstock for our plastics or into the fuels we run our cars. However, there are risks associated with relying on such a singular substance as the cornerstone of modern society. What happens to the atmosphere when we burn too much fuel? Will we ever run out? How are other species affected by the drilling, processing, and transportation of petroleum? Why do human political conflicts in one region affect energy scarcity in *another*? These are all questions with sundry answers. What if there were other ways to manufacture those same chemicals, different formulations for our medications, or compatible fuels for our vehicles? Remarkably, the world around us is filled with innovation. Bacteria, yeast, plants, animals, even humans contain complex and highly efficient machines that can create complex materials that rival those made of traditional energy sources. These machines are a product of biochemistry, a mixture of DNA and proteins and the pieces in-between that make life on this planet possible. With knowledge of these systems and their intricacies, we set out to harness their potential to satisfy our demands. Herein, novel techniques are demonstrated that (i) generate better understanding of these organisms at the cellular and genomic levels and (ii) enable engineering of their metabolism to enhance production of valuable chemicals and biofuels.

Non-renewable resources, such as petroleum, coal, and natural gas have been marred by price instabilities and environmental backlash, creating incentives in other markets to fill the

void. Hydro, solar, wind, and other diverse technologies which harvest energy from waves and thermal vents are seeing rapid ascension in utilization [1]. Though they do create the energy to power cities and factories, they produce poor transportation fuel and cannot be used to create organic compounds. Common chemicals like acetic acid, ethylene, xylene, ethanol, and thousands of others are created from the simple petroleum base product [2]. From the simpler chemicals, more complex ones can be created. Polyamide, or common nylon, for example, can be manufactured from caprolactam which is synthesized from cyclohexanone, which in turn comes from cyclohexane which derives from benzene, a common fraction of crude oil [3]. However, micro-organisms have been producing some of these materials for millions of years, and they have likely generated thousands more which have gone undetected to date [4]. Application of microorganisms as "microbial cell factories" (MCFs) stretches from ancient alcohol fermentation to industrial acetone, butanol, and ethanol (ABE) fermentation [5] to the modern production of flavonoids and "anti-aging" nutraceuticals [6].

Bioprocessing, or the industrial processing of biological materials, has many advantages over traditional chemical processing. One of these advantages is the potential for net zero carbon. Fossil fuels rely on fossils, or organisms that existed hundreds of millions of years ago, taking up carbon and carbon dioxide from the environment, which were slowly compressed into a crude oil deep below the Earth's crust. As it is unlikely we have the liberty of hundreds of millions of years for this process to repeat itself, bioprocessing aims to make environmental carbon readily available in chemical form. MCF's (i.e., organisms engineered for the optimal production of native or non-native compounds) often get their energy from simple sugars like glucose or, ideally, complex polymers of sugars like cellulose. The origin of both of these sugars must be autotrophic, or an organism capable of extracting energy from sunlight. This solar energy

enables the autotroph to convert H_2O into oxygen and chemical energy. This energy enables the enzyme ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) to fix CO_2 in the atmosphere into a source usable by other organisms. In a perfectly closed system, sugars containing fixed CO_2 are consumed to generate combustible fuels or consumable foods, which undergo combustion or catabolism and release CO_2 back into the atmosphere, forming a net zero change in carbon.

Reducing CO_2 release into the environment is critical to reducing or postponing the effects of global climate change [7]. Energy use alone accounts for 57% of worldwide greenhouse gas emissions [8] and 32% of US emissions. Combustion of gasoline in vehicles is responsible for 28% of US CO_2 emissions [9] These numbers could be reduced or eliminated if fuels were produced in a manner (i.e. sustainably) such that an equivalent amount of CO_2 was produced to generate the fuel as consuming it. Researchers are divided as to whether biofuels such as ethanol will be a solution due to the massive land requirements and the transport, crop fertilization, and fermentation requirements. Some estimates yield a 29-118% net loss in energy depending on ethanol or biodiesel produced, available infrastructure, or whether cellulosic substrates are involved [10]. Recent reviews summarizing available technologies also appear to agree that none of the implemented technologies for biofuels can release more energy than the sum of the inputs, but they do yield savings of 50-78% in total greenhouse gas emissions than conventional fuels [11]. This is not to say that newer pipeline technologies may become energy neutral. Research groups have looked at using plants like cyanobacteria or CO_2 fixing *Clostridium ljungdahlii* to convert CO_2 directly into ethanol, acetate and hydrocarbons [12, 13], which have the potential to drastically reduce the carbon footprint of the processes if developed.

Use of microbial cell factories to produce chemicals

Chemical production makes up only 13% of petroleum demand—less than half that of transportation fuels—and has become an attractive target for replacement by bio-based chemicals by the US Department of Energy [14, 15]. Even if we disregard lifecycle carbon and energy balances, some chemicals can be produced in fewer steps using biological systems than chemical synthesis. Additionally, these reactions can be done so at physiological temperatures and pressures and often without generating solvent waste associated with traditional petrochemical processes. For example, the production of 1,3 propanediol (1,3PDL) is necessary for polytrimethylene terephthalate (PTT) which has seen surge in demand due to its unique structural properties in the textile industry. DuPont (Wilmington, DE) has developed technology to create 1,3PDL from agricultural glucose and glycerol from animal products using fermentation by *Clostridium butyricum* modified with foreign genes. Not only does this fermentation method avoid the toxic and expensive processing steps in the petroleum based synthesis, it also produces higher yields and is more strongly poised to satisfy the global demand [16]. Another example includes the production of phenol from glucose via shikimic acid using modified *E. coli* harboring a plasmid with foreign *glk*, *tktA*, and *glf* genes instead of the carcinogenic and petroleum based chemical benzene as the precursor [17].

Advances in metabolic engineering to common barriers

Metabolic engineering remains far from a static science. Advances in manufacturing, DNA synthesis, protein discovery, information processing, and sample discovery have led to instrumental increases in productivity in biological synthesis. Some common barriers to production include: (i) accumulation of toxic products or intermediates, (ii) slow growth or

survivability rates of the organism, (iii) poor enzyme kinetics or specificity, (iv) high cost of feed sources and (v) poor integration of foreign pathways within a host. A brief explanation of these cases is given below along with particular examples where researchers have overcome them.

i. The most studied products of metabolic engineering to date have been alcohols due to their use as a transportation fuel and compatibility with the current gasoline infrastructure. Alcohols traditionally serve as the end product of fermentation, allowing cells to recycle NADH or NADPH. However, the alcohols are toxic to cells, causing changes in membrane structure, lipid composition, oxidative stress and cell behavior. Some of the most recent engineering approaches to improve upon cellular tolerance involve modifying the cell membrane by including sterols or saturated fatty acids in the membrane [18], designing transporter proteins to export alcohol outside the cell [19, 20], or adding quencher proteins to absorb free oxygen radicals before they can inhibit cell growth [21]. In one specific case, an alcohol dehydrogenase from *Clostridium thermocellum* ethanol tolerant mutant was computationally simulated and indicated that the cofactor specificity changed from NADH-dependent to NADPH-dependent. The authors claimed that cofactor changes re-directed electron flow [22].

ii. Novel cell or enzyme function does not always integrate well into industrial settings. One example is with *Clostridium cellulolyticum*, an anaerobe capable of breaking down cellulose, hemicellulose, and some lignins without pre-processing. One caveat is that it suffers from metabolic bottlenecking where, under carbon excess (such as in a lab or production setting) pyruvate builds up inside the cell to inhibitory concentrations, resulting in an inefficient growth rate [23]. Once such solution was the overexpression of pyruvate consuming enzymes

from alternative organisms to alleviate the buildup of pyruvate yielded an increase in the rate of cellulose consumption by 150% [24].

iii. Enzymes are the catalytic workhorses of cells, responsible for reducing the timescale at which chemicals interact and orient themselves to form bonds. Often, the function of a single enzyme in a pathway is a bottleneck where the structure of the enzyme cannot biophysically function at a faster rate. Advances in cloning and gene modification have yielded many examples of industrial enzymes with optimized reaction velocity, temperature stability, and cofactor usage [25]. Mutating enzymes to utilize chemicals with small changes in chain length or additional side groups from the native substrate has found utility when no native acting enzyme has been documented, creating novel compounds or particular enantiomer product ratios [26, 27].

iv. One of the largest energy requirements and sources of CO_2 emissions for microbial chemical production comes from the processing of agricultural materials for glucose [11]. Instead of utilizing corn and soy crops with high fertilizer requirements, agricultural wastes could be used instead. The most dominant being lignocellulose, which can be converted into simple sugars. However, the complex matrix of celluloses, hemicellulose, and lignin requires expensive enzymatic, chemical, or mechanical extraction which can complicate downstream processing [28]. A consolidated bioprocessing approach involves organisms capable of growing on lignocellulose for the production of desired chemicals. While strains including *Clostridium cellulolyticum* and *Clostridium thermocellum* are often candidates for these systems [29], metabolic engineering of the cellulolytic pathways into yeast [30] and *E. coli* [31] have been attempted.

v. Poor integration of foreign genes or overexpression of native genes in an MCF can lead to competition for shared resources, sacrificing growth and cell health for production the desired product. One example is the production of flavonoids. The early integration of the three genes for flavanones synthesis in *E. coli* resulted in a depletion of malonyl-CoA and ATP. Computational constraint based models predicted gene knockouts to re-route carbon to the necessary pathways for optimal flavonoid [32] and succinic acid production [33].

The difference between rational and combinatorial metabolic engineering

Two methods of overcoming the barriers to biological production are (i) rational and (ii) combinatorial metabolic engineering designs. Rational designs include all "knowledge-based" strategies (e.g., knockout of a gene to disrupt a competing pathway). This includes construction of a "minimal cell" by knocking out all unnecessary genes in *E. coli*. This has allowed researchers to maximize the flux of glucose and xylose to biomass and ethanol without formation of secondary products [34]. Less drastic targeted approaches include the work done with *C. cellulolyticum* to resolve the over-accumulation of pyruvate by expression of enzymes capable of degrading pyruvate at a faster rate [24]. Combinatorial designs, on the other hand, start with the admission that knowledge is limited for a desired trait (e.g., no single gene exists that can make cells tolerate larger concentrations of alcohol). Combinatorial designs take advantage of large mutational pools and enrichment screens to identify beneficial adaptations. While the rational design approach has been used successfully in conjunction with metabolic models and metabolic modeling programs like OptKnock [35], OptForce [36], and FBrAtio [37], the combinatorial approach has potential for biological discovery while yielding cultures with desired traits. Additionally, the rapid decrease in the cost of DNA sequencing makes many of the shotgun

designs more feasible because of the ability to test analyze the entire sample space. The focus of this dissertation, as will be seen, is the application of just such methods, novel means of generating the libraries of random DNA fragments, their inspection with next-generation high throughput sequencing technologies, their high-throughput cloning and expression in bacteria, and the multiple enrichment strategies that identify phenotypes with remarkable improvements in cellular fitness.

Deployment of genomic DNA libraries

Genomic DNA libraries were originally constructed before the age of PCR. Packaging enzymatically digested DNA into a library allowed for isolation of a gene unit. A "DNA library" is a collection of plasmids each of which has a different fragment of DNA cloned into a specific position of the plasmid. The mixture is very heterogeneous and is designed to encompass DNA fragments that represent an entire genome. Further digesting DNA libraries enabled production of restriction maps for a genome [38]. Cells were often mutated with chemicals like N′-nitro-N-nitrosoguanidine [39] or ultra-violate radiation [40] to induce loss or gain of function which can be detected by growth on selective media. Generation of a screening library from the selected mutants allowed for finding the most basic unit responsible for the phenotype. Viruses and transposon based mutation methods have also been developed, swapping or recombining regions of the host genome [41], and have seen some use in hard-to-transform organisms where a single transposon bearing mutant could propagate multiple mutations in a culture.

A few examples of genomic DNA libraries used in metabolic engineering include a search for DNA fragments imparting 1-butanol tolerance in *C. acetobutylicum*. In this study, DNA from *Clostridium* was sheared down to 1-5kb using sonication and packaged into a shuttle

vector capable of propagating in *E. coli* and the native *C. acetobutylicum*. After exposure to successively higher amounts of 1-butanol, samples were taken and hybridized to microarrays containing all genes in *C. acetobutylicum*. The microarrays suggested that successive generations of butanol stress refined the number and diversity of enriched fragments from the full library until isolation of genes capable of providing up to 81% greater tolerance to butanol were developed [42].

Another method the depends heavily on genomic library construction is the multiSCale Analysis of Library Enrichment (SCALEs) which uses DNA libraries of different sized fragments between 500 and 8000 bp generated by restriction digestion of *E. coli* genomic DNA and ligation to a commercial plasmid. Microarrays were used to obtain hybridization snapshots of the state of the library at each enrichment point. Each enriched fragment was scored based on abundance and fragment length to rank the effect of given clusters enriched in cells exposed to a stress [43]. This technique has been used to identify genes that are responsible for sustaining cellulosic hydrolysate toxicity [44], acetate tolerance [45], overcoming anti-metabolites [46], ethanol tolerance [47] and ethanol production [48]. While many genes may have synergistic or tandem effects, the way at which multiple, long distant genes interact was also investigated. Two libraries of 3-6 kb genomic library fragments were co-expressed in *E. coli*, along with a fosmid library containing 35 kb sized inserts to find advances in acid tolerance of up to 9000-fold [49].

Outline of experiments and findings

The work presented in this dissertation started with the aims of generating a DNA library to enhance the growth rate of *C. cellulolyticum* when grown on cellulosic substrates. Enriched genes would aid in filling metabolic bottlenecks and ultimately optimize the organism further for

production of cellulosic chemicals and biofuels. However, attempts to follow published methods of generating DNA libraries led to inefficiencies and DNA libraries that were very small in size (i.e., they contained few different DNA fragments). The original methodology of DNA library construction used high pressure nebulization to mechanically shear bulk DNA. This then required a purification step and an end-repair step to remove any mismatched DNA tails. Next, the blunted DNA was phosphorylated and combined with a dephosphorylated plasmid, linearized with a blunt-cutting enzyme [50]. A large amount of variation can take place between the above steps. DNA that makes up the library can be digested with blunt cutter enzymes or mechanically sheared. In another approach, once purified and size separated by agarose electrophoresis, the DNA is end-repaired, dephosphorylated, and adenine ("A") tailed using *Taq* polymerase. A-Tailed DNA can be ligated into a commercial T/A cloning vector directly [42], while blunt DNA can take additional routes: (i) cloning into commercial blunt-ended vector [43], (ii) ligated to a blunt digested vector [51], or (iii) ligated into a PCR amplified linear plasmid [46] (Figure 1-1). Ligated DNA is then transformed into a variety of commercial cells by either chemical transformation or electroporation and transformants are grown out into the first generation of the enriching genomic DNA library. Due to the inefficiencies experienced with end-polishing DNA and blunt-end cloning (and even T/A cloning) a new method of DNA library construction was explored. Thus, we explored PCR for the generation of DNA fragments to be cloned to produce the DNA library. The method, herein, uses a PCR amplification technique, which can remove many of the processing steps between the host genome and the final library. The method is based on degenerate oligonucleotide primed PCR (DOP-PCR). The technique relies on PCR primers with degenerate regions (i.e., bases with equal probability of being A/T/G/C). These primers then bind randomly to genomic DNA and amplify different fragments during a PCR run. If the *Taq*

polymerase is used with this procedure, the resulting DNA fragments will be A-tailed, and we experienced very high efficiency of cloning these DNA fragments into commercial vectors. The procedure allowed for construction of high-quality libraries and provided other benefits that are not possible with traditional methods of construction such as: (i) the ability to produce libraries from ng quantities of DNA, (ii) the ease of producing libraries from non-culturable metagenomes, and (iii) the ability to purposely induce "errors" in DNA sequences to expand genomic diversity. As the DOP-PCR approach is a transgression from traditional techniques, we used next-generation high-throughput sequencing to verify the quality of the DNA libraries.

The organization of this dissertation is as follows. The results of DNA library construction by DOP-PCR are discussed in Chapter 2. It is demonstrated that the DOP-PCR method can represent up to 95% of the host genome. In addition, DOP-PCR primers can be designed based on the laws of thermodynamics to reduce the bias these primers may have for amplifying one region of a genome over another. The DNA libraries constructed using DOP-PCR were used in Chapter 3 to seek DNA fragments amplified from *E. coli*, *C. acetobutylicum*, and a mixed environmental soil metagenome that impart a tolerance advantage on *E. coli* growing under exposure to 1-butanol. Once DNA fragments were isolated, the resulting strains were analyzed with Raman spectroscopy to understand the cell physiology changes produced by those DNA fragments that led to improved tolerance. The approach in Chapter 4 used a similar method to search for overexpression candidates in the *C. cellulolyticum* and the metagenomic DNA library to overcome the metabolic burdens associated with cellulosic substrate degradation. This project identified a single 350 bp DNA fragment capable of improving the yield of ethanol and pyruvic acid from cellobiose by 250% compared to the control. Chapter 5 focuses on our development of a synthetic gene circuit that was designed to sense the concentration of L-

arginine in a cell and activate a cellular toxin when L-arginine concentrations are low. The circuit is called the Feedback Inhibition of Transcription for Substrain Selection (FITSelect). When used with a combinatorial metabolic engineering strategy, this approach ties the production of L-arginine to the growth rate of an individual cell. Only those cells receiving a mutation or DNA fragment making it capable of higher L-arginine production will survive the action of the circuit. This chapter describes the construction, testing, and validation of the circuit. The circuit was then used for metabolic engineering by transforming a DNA library in effort to select for DNA fragments resulting in increased L-arginine production.

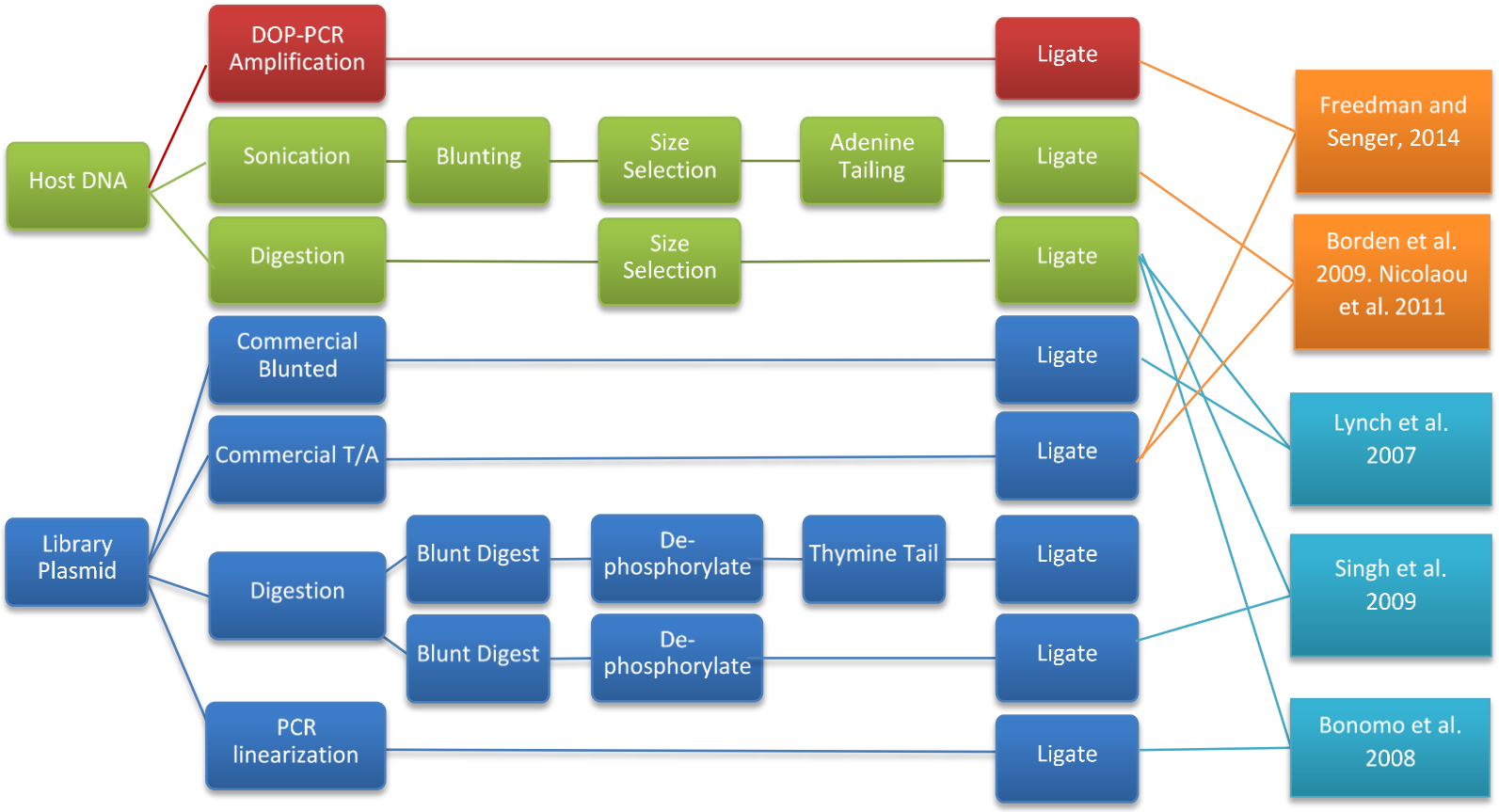


Figure 1-1. Summary of techniques used to generate genomic DNA libraries for strain enrichment. Successfully used for host DNA preparation (green), plasmid preparation (blue), and the methods of this dissertation (red). Whether studies used electroporation (turquoise) or chemical transformations (orange) is also indicated.

References

[1] Twidell, J., Weir, A. D., *Renewable energy resources*, Taylor & Francis 2006.

[2] Matar, S., Hatch, L. F., *Chemistry of petrochemical processes*, Gulf Professional Publishing 2001.

[3] Trossarelli, L., The history of nylon. *Commissione Centrale Materiali e Tecniche* 2003.

[4] Tiedje, J. M., Asuming-Brempong, S., Nüsslein, K., Marsh, T. L., Flynn, S. J., Opening the black box of soil microbial diversity. *Applied Soil Ecology* 1999, *13*, 109-122.

[5] Katzen, R., Tsao, G., A View of the History of Biochemical Engineering, in: Fiechter, A. (Ed.), *History of Modern Biotechnology II*, Springer Berlin Heidelberg 2000, pp. 77-91.

[6] Leonard, E., Lim, K.-H., Saw, P.-N., Koffas, M. A., Engineering central metabolic pathways for high-level flavonoid production in Escherichia coli. *Applied and environmental microbiology* 2007, *73*, 3877-3886.

[7] Cramer, W., Bondeau, A., Woodward, F. I., Prentice, I. C., *et al.*, Global response of terrestrial ecosystem structure and function to CO2 and climate change: results from six dynamic global vegetation models. *Global Change Biology* 2001, *7*, 357-373.

[8] Scheinbaum, C., Schipper, L., unpublished article for US Department of Energy (DOE) and the Environmental Protection Agency (EPA) 1993.

[9] Environmental Protection Agency (2014). Inventory of US Greenhouse Gas Emissions and Sinks: 1990-2012. Washington, D.C.

[10] Pimentel, D., Patzek, T., Ethanol Production Using Corn, Switchgrass, and Wood; Biodiesel Production Using Soybean and Sunflower. *Nat Resour Res* 2005, *14*, 65-76.

[11] Larson, E. D., A review of life-cycle analysis studies on liquid biofuel systems for the transport sector. *Energy for Sustainable Development* 2006, *10*, 109-126.

[12] Leang, C., Ueki, T., Nevin, K. P., Lovley, D. R., A Genetic System for Clostridium ljungdahlii: a Chassis for Autotrophic Production of Biocommodities and a Model Homoacetogen. *Applied and Environmental Microbiology* 2013, *79*, 1102-1109.

[13] Tan, X., Yao, L., Gao, Q., Wang, W., *et al.*, Photosynthesis driven conversion of carbon dioxide to fatty alcohols and hydrocarbons in cyanobacteria. *Metabolic engineering* 2011, *13*, 169-176.

[14] Dodds, D. R., Gross, R. A., Chemicals from biomass. *Science* 2007, *318*, 1250-1251.

[15] Werpy, T., Petersen, G., Aden, A., al., e., in: Energy, U. S. D. o. (Ed.) 2004.

[16] Saxena, R. K., Anand, P., Saran, S., Isar, J., Microbial production of 1,3-propanediol: Recent developments and emerging opportunities. *Biotechnology Advances* 2009, *27*, 895-913.

[17] Gibson, J. M., Thomas, P. S., Thomas, J. D., Barker, J. L., *et al.*, Benzene-Free Synthesis of Phenol. *Angewandte Chemie* 2001, *113*, 1999-2002.

[18] Caspeta, L., Chen, Y., Ghiaci, P., Feizi, A., *et al.*, Altered sterol composition renders yeast thermotolerant. *Science* 2014, *346*, 75-78.

[19] Yang, K.-M., Woo, J.-M., Lee, S.-M., Park, J.-B., Improving ethanol tolerance of Saccharomyces cerevisiae by overexpressing an ATP-binding cassette efflux pump. *Chem Eng Sci* 2013, *103*, 74-78.

[20] Dunlop, M. J., Dossani, Z. Y., Szmidt, H. L., Chu, H. C., *et al.*, Engineering microbial biofuel tolerance and export using efflux pumps. *Molecular systems biology* 2011, *7*.

[21] Chin, W. C., Lin, K. H., Chang, J. J., Huang, C. C., Improvement of n-butanol tolerance in Escherichia coli by membrane-targeted tilapia metallothionein. *Biotechnology for biofuels* 2013, *6*, 130.

[22] Brown, S. D., Guss, A. M., Karpinets, T. V., Parks, J. M., *et al.*, Mutant alcohol dehydrogenase leads to improved ethanol tolerance in Clostridium thermocellum. *Proceedings of the National Academy of Sciences* 2011, *108*, 13752-13757.

[23] Payot, S., Guedon, E., Cailliez, C., Gelhaye, E., Petitdemange, H., Metabolism of cellobiose by Clostridium cellulolyticum growing in continuous culture: evidence for decreased NADH reoxidation as a factor limiting growth. *Microbiology* 1998, *144*, 375-384.

[24] Guedon, E., Desvaux, M., Petitdemange, H., Improvement of Cellulolytic Properties of Clostridium cellulolyticum by Metabolic Engineering. *Applied and Environmental Microbiology* 2002, *68*, 53-58.

[25] Fisher, A. K., Freedman, B. G., Bevan, D. R., Senger, R. S., A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories. *Computational and Structural Biotechnology Journal* 2014, *11*, 91-99.

[26] Cahoon, E. B., Lindqvist, Y., Schneider, G., Shanklin, J., Redesign of soluble fatty acid desaturases from plants for altered substrate specificity and double bond position. *Proceedings of the National Academy of Sciences* 1997, *94*, 4872-4877.

[27] Pazmiño, D. E. T., Snajdrova, R., Rial, D. V., Mihovilovic, M. D., Fraaije, M. W., Altering the Substrate Specificity and Enantioselectivity of Phenylacetone Monooxygenase by Structure-Inspired Enzyme Redesign. *Advanced Synthesis & Catalysis* 2007, *349*, 1361-1368.

[28] Sun, Y., Cheng, J., Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource Technology* 2002, *83*, 1-11.

[29] Johnson, E. A., Sakajoh, M., Halliwell, G., Madia, A., Demain, A. L., Saccharification of complex cellulosic substrates by the cellulase system from Clostridium thermocellum. *Applied and environmental microbiology* 1982, *43*, 1125-1132.

[30] Chang, J.-J., Ho, C.-Y., Ho, F.-J., Tsai, T.-Y., *et al.*, PGASO: A synthetic biology tool for engineering a cellulolytic yeast. *Bitechnol Biofuels* 2012, *5*, 53.

[31] Srivastava, R., Kumar, G. P., Srivastava, K. K., Construction of a recombinant cellulolytic Escherichia coli. *Gene* 1995, *164*, 185-186.

[32] Fowler, Z. L., Gikandi, W. W., Koffas, M. A. G., Increased Malonyl Coenzyme A Biosynthesis by Tuning the Escherichia coli Metabolic Network and Its Application to Flavanone Production. *Applied and Environmental Microbiology* 2009, *75*, 5831-5839.

[33] Lee, S. J., Lee, D.-Y., Kim, T. Y., Kim, B. H., *et al.*, Metabolic Engineering of Escherichia coli for Enhanced Production of Succinic Acid, Based on Genome Comparison and In Silico Gene Knockout Simulation. *Applied and Environmental Microbiology* 2005, *71*, 7880-7887.

[34] Trinh, C. T., Unrean, P., Srienc, F., Minimal Escherichia coli Cell for the Most Efficient Production of Ethanol from Hexoses and Pentoses. *Applied and Environmental Microbiology* 2008, *74*, 3634-3643.

[35] Burgard, A. P., Pharkya, P., Maranas, C. D., Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering* 2003, *84*, 647-657.

[36] Ranganathan, S., Suthers, P. F., Maranas, C. D., OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* 2010, *6*, e1000744.

[37] McAnulty, M. J., Yen, J. Y., Freedman, B. G., Senger, R. S., Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC systems biology* 2012, *6*, 42.

[38] Kohara, Y., Akiyama, K., Isono, K., The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell, 50*, 495-508.

[39] Mandell, J. D., Greenberg, J., A new chemical mutagen for bacteria, 1-methyl-3-nitro-1-nitrosoguanidine. *Biochemical and Biophysical Research Communications* 1960, *3*, 575-577.

[40] Radman, M., SOS Repair Hypothesis: Phenomenology of an Inducible DNA Repair Which is Accompanied by Mutagenesis, in: Hanawalt, P., Setlow, R. (Eds.), *Molecular Mechanisms for Repair of DNA*, Springer US 1975, pp. 355-367.

[41] Simon, R., Priefer, U., Pühler, A., A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat Biotechnol* 1983, *1*, 784-791.

[42] Borden, J. R., Papoutsakis, E. T., Dynamics of genomic-library enrichment and identification of solvent tolerance genes for Clostridium acetobutylicum. *Applied and environmental microbiology* 2007, *73*, 3061-3068.

[43] Lynch, M. D., Warnecke, T., Gill, R. T., SCALEs: multiscale analysis of library enrichment. *Nat Meth* 2007, *4*, 87-93.

[44] Mills, T. Y., Sandoval, N. R., Gill, R. T., Cellulosic hydrolysate toxicity and tolerance mechanisms in Escherichia coli. *Biotechnology for biofuels* 2009, *2*, 26.

[45] Sandoval, N. R., Mills, T. Y., Zhang, M., Gill, R. T., Elucidating acetate tolerance in *E. coli* using a genome-wide approach. *Metabolic engineering* 2011, *13*, 214-224.

[46] Bonomo, J., Lynch, M. D., Warnecke, T., Price, J. V., Gill, R. T., Genome-scale analysis of anti-metabolite directed strain engineering. *Metabolic Engineering* 2008, *10*, 109-120.

[47] Woodruff, L., Pandhal, J., Ow, S. Y., Karimpour-Fard, A., *et al.*, Genome-scale identification and characterization of ethanol tolerance genes in< i> Escherichia coli</i>. *Metabolic engineering* 2013, *15*, 124-133.

[48] Woodruff, L., Boyle, N. R., Gill, R. T., Engineering improved ethanol production in *Escherichia coli* with a genome-wide approach. *Metabolic engineering* 2013, *17*, 1-11.

[49] Nicolaou, S. A., Gaida, S. M., Papoutsakis, E. T., Coexisting/Coexpressing Genomic Libraries (CoGeL) identify interactions among distantly located genetic loci for developing complex microbial phenotypes. *Nucleic Acids Research* 2011, *39*, e152.

[50] Sambrook, J., Russell, D., Irwin, N., Molecular cloning: a laboratory manual-/Joseph Sambrook, David W. Russell;[assoc. ed.: Nina Irwin, Kaaren A. Janssen]. 2001.

[51] Singh, A., Lynch, M. D., Gill, R. T., Genes restoring redox balance in fermentation-deficient E. coli NZN111. *Metabolic Engineering* 2009, *11*, 347-354.

Chapter 2

Thermodynamic design of degenerate oligonucleotide primers for whole genome DNA amplification with reduced bias for use in DNA libraries

ABSTRACT

The production of genomic DNA libraries for metabolic engineering and biological discovery applications has been simplified through the use of degenerate oligonucleotide primed PCR (DOP-PCR). This procedure can perform whole genome amplification and result in gene-size DNA fragments that clone with very high efficiency into commercial linearized T-tailed vectors. This research focused on the design of primers for DOP-PCR that can produce libraries of adequate DNA fragment size while maintaining full genome coverage and minimizing bias towards amplifying only certain regions of a genome. The basic primer design consists of a static 5' region, followed by a guanine ("G") rich region (6-9 residues), and a degenerate (equal probability of A/T/G/C) region of 3-6 residues. Thermodynamics were employed to design primer pools so each combination of degenerate region could exist without being confined by a hairpin structure. Primers were tested by DOP-PCR amplification of 3 genomes: *E. coli* NEB 10-beta (50.8% GC content), *Clostridium acetobutylicum* ATCC 824 (30.9% GC), and *Comamonas testosteroni* ATCC 11996 (61.8% GC). Results from 15 primer candidates and all genomes were analyzed by next-generation sequencing. Results showed significant improvement of the thermodynamically designed primers to a published primer that served as a benchmark for this

study. Pooling of thermodynamically designed primers led to 84-97% genome coverage, DNA fragment sizes between 1-6 kb, and reduced amplification bias and error relative to the benchmark primer. Furthermore, the addition of 70 μM $MnCl_2$ to the DOP-PCR yielded "error-prone" DNA libraries that could hold tremendous potential for metabolic engineering applications.

INTRODUCTION

Applications of genomic DNA libraries

The genomic DNA library has become an important tool in elucidating genetic structure, probing disease loci, identifying protein interactions, whole genome sequencing, and mining metagenomes for novel enzyme activity [1-4]. A DNA library is a heterogeneous mixture of cloned DNA fragments that often originate from fractionated genomic (or metagenomic) DNA. In the face of selective pressure (i.e., stress), A DNA library often works to increase cellular fitness, as those cells containing DNA fragments that offer a growth advantage ultimately dominate the culture. For example expression libraries tested on auxotrophic microbes which lack the ability to produce vital amino acids through a single [5] or multi-gene [6] knockouts were rescued byexpressing appropriately designed DNA libraries. Wide-scope transcriptomics using DNA microarrays have identified fragments from genomic DNA libraries that become enriched or depleted as a result of alcohol stress [7, 8], cell growth inhibitors [9, 10], xenobiotics [11], oxidative or heat stress [12] or natural cell processes including growth phase transitions or sporulation [13]. Genomic DNA libraries can also be constructed to collect and propagate DNA from un-culturable microbes or from samples that exist in limited quantities to reveal a wealth of

information relating to microbial diversity [14, 15], human health [16], or criminal forensics [17]. DNA of organisms from exotic or extreme environments can also be packaged into a genomic DNA library and screened for enzymes with novel catalytic functions, such as polyester and polyurethane degradation by a fungi isolated from the Ecuadorian rainforest [18].

DNA library construction methods

Genomic DNA libraries must satisfy two broad metrics: (i) genome coverage and (ii) avoidance of bias. A genomic DNA library that does not contain all of the information of the host genome may lead to sub-optimal results and could completely miss a gene of interest that has a desired catalytic action. A highly biased sample contains a disproportionate number of certain DNA fragments. This skews the enrichment process and may impact screening, where a DNA fragment appears enriched due to its large starting concentration rather than its enrichment through the ability to confer fitness. The traditional approach to genomic DNA library construction and cloning is first fragmentation of purified genomic DNA through mechanical forces or enzymatic digestion. While restriction enzymatic digestion can create overhangs in DNA that simplify later cloning, they also only cut at defined locations, leading to significant bias. Mechanical forces shear DNA with less discernable bias [19], but the resulting DNA fragments require end-polishing to enable effective cloning into a plasmid [20].

Over the course of our research, we sought a reliable method for constructing genomic DNA libraries of 1-10 kb DNA fragments with high cloning efficiencies from the genomic DNA of difficult to cultivate microbes. The traditional approach of mechanical shearing of genomic DNA, end-polishing, and blunt-end cloning yielded poor results (i.e., poor colony counts with effectively cloned DNA fragment inserts) over a long period of time. Alternative methods of

DNA library construction were sought, leading to the investigation of whole genome DNA PCR-based amplification. This strategy is based on the premise of perfectly random priming of a DNA template and unbiased amplification by the polymerase in a PCR run. There are typically three types of whole genome amplification by PCR: (i) degenerate oligonucleotide primed PCR (DOP-PCR), (ii) primer extension pre-amplification PCR (PEP-PCR), and (iii) the non-PCR-based multiple displacement amplification (MDA). These methods are used to amplify trace amounts of DNA, often to the picogram level present in a single cell to perform medical or forensic diagnostics [21]. The amplification methods have also been used in characterization of metagenomic samples and sequencing of mixed populations of soil microbes [22-25]. The DOP-PCR technique originally involved the primer DOPcontrol (5′-CCGACTCGAGNNNNNNATGTGG-3′) to generate fragments in the range of 500-2000 nucleotides using a two-step PCR program of low and high stringency annealing with a non-displacing polymerase. The 5' end was designed to contain an XhoI restriction site and the degenerate "N" region was meant to randomize binding across sequences ending in "ATGTGG" [26]. The PEP-PCR method is similar to DOP-PCR, using instead a random 15-mer (5'-NNN-NNN-NNN-NNN-NNN-3') and up to 50 PCR cycles to amplify target DNA with a polymerase enzyme. The MDA technique uses a highly processive strand displacing polymerase and random hexamer oligonucleotides to amplify large (>15 kb) fragments under isothermal conditions [27].

Determining the efficiency of whole genome amplifications has relied on identifying a set number (usually 20 to 50 per genome) of pre-determined sequences or microsatellites [28, 29], FISH hybridization [30], or Multiplex STR amplification [31]. Microsatellite sampling techniques have revealed that the three methods perform whole genome amplification to differing degrees. PEP-PCR using a 15N-mer had a 78% probability of amplifying a whole

genome from a single cell at least 30 times [32]. DOP-PCR studies indicated 200-600 fold amplification of a human size genome allowing for highly accurate microsatellite detection [33], while others studies counter this finding to show the method may preferentially amplify some microsatellites 10^6-fold more than others. The same research calculated that MDA, in addition to producing longer fragments, only produced a 3-fold bias between markers [34].

Comparisons of methods for whole genome amplification

The increased availability of next-generation sequencing technologies has led to an in-depth examination of the amplification technologies at single-base pair resolution. Comparison of the coverage and distribution of PEP-PCR, DOP-PCR, and MDA were assessed by this technology and comparison of amplified and unamplified genomes was performed using the Kolomogorov-Smirnov statistical test of the reconstructed and aligned genomes. Results concluded that the best method of whole genome amplification was the MDA method with a phi 32 strand displacing polymerase due to its fidelity and random binding. In this case, DOP-PCR failed to represent more than 9.7% of a *Halobacterium* genome [35]. A more recent approach examined amplification of the *Caenorhabditis elegans* genome using different MDA protocols and summarized the variations in the Illumina® Next Generation sequencing reads compared to an unamplified sample, confirming that these methods approach full genome coverage yet still maintain a distinct amplification bias under all formulations [36].

Despite its reported short-comings, DOP-PCR still holds promise due to (i) the ability to produce gene-size fragments using widely available polymerases, (ii) the tendency of polymerases to insert a 3' adenine ("A") residue on the terminal ends of PCR products, and (iii) the ability to incorporate a defined barcode sequence on the 5' end to facilitate downstream

cloning and sequencing [22]. Efforts have been made to determine the effects of PCR cycling parameters, choice of polymerase, initial DNA template concentrations, and primer design. Design modifications include altered cycling parameters to generate fragments up to 10 kb in length [28], monitoring of low temperature, non-stringent PCR cycles to promote random priming, determining the advantages of the Deep Vent$_R$™ proofreading polymerase [31], primer design containing 6 random degenerate ("N," equal probability of A/T/G/C) nucleotides, 5-6 base pairs providing an anchor on the 3' end, and an optimal primer concentration of 5 μM [30]. An additional study suggested 6 degenerate ("N") base pairs in the middle of the primer along with a 5-6 bp region of 50% GC content at the terminal 3' end yields the most sensitive and consistent DOP-PCR amplification [37].

Thermodynamic design of DOP-PCR primers

In this research, we developed a thermodynamics-based method for designing DOP-PCR primers for whole genome amplification that strikes a balance between genome coverage and bias. The method also has utility that extends far beyond that of a single genome, allowing production of gene-sized fragments from metagenomic samples and trace amounts of DNA from virtually any source. The method is of significance because the A-tailed DNA products of DOP-PCR have been found to clone very easily and consistently to produce DNA libraries. We have used nucleotide binding and folding thermodynamic calculations to design DOP-PCR primer sequences that cover a maximum annealing space by minimizing self-priming and hairpin-prone sequences. Next-generation sequencing of the resulting amplified DNA libraries has confirmed the newly designed primers provide distinct advantages over previous designs, leading to more complete genomic coverage and reduced bias. The newly designed primers proved applicable to

microbes of widely varying GC content, both Gram positive and negative microbes, those containing megaplasmids, and metagenomic DNA. In our experience, this method of DOP-PCR library generation, cloning, transformation, and expression can be executed in about 1 week by graduate and undergraduate researchers with moderate molecular biology experience. Importantly, the new methodology is highly repeatable and reliable for constructing DNA libraries.

MATERIALS AND METHODS

Strains and cultures

Cultures of *E. coli* NEB 10-beta (a K12 DH10B variant) (New England Biolabs; Ipswich, MA), *Clostridium acetobutylicum* ATCC 824, and *Comamonas testosteroni* ATCC 11996 were used in this research. The strains were chosen for this research based on their genomic GC content. *E. coli* has neutral genomic GC content (50.8%), *C. acetobutylicum* has low genomic GC content (30.9%), and *C. testosteroni* has high genomic GC content (61.8%). All cultures were maintained in glycerol stocks at -80°C. *E. coli* and *C. testosteroni* were grown at 37ºC in Lysogeny Broth (LB) medium. *C. acetobutylicum* was grown at 37ºC in an $N_2/CO_2/H_2$ (90/5/5%) environment in a Bactron anaerobic chamber (Shel Lab; Cornelius, OR) on 2xYTG medium [38] without antibiotics. All cultures were grown initially on agar plates, and experiments were carried out with cultures grown to late exponential phase in liquid medium. DNA was extracted from *E. coli* and *C. testosteroni* cultures using Generation Capture columns (Qiagen; Valencia, CA) according to manufacturer's directions. DNA was extracted from *C. acetobutylicum* as previously described [39]. Culture density in liquid media was monitored by optical density at 600 nm (OD_{600}).

DOP-PCR DNA amplification

Genomic DNA concentrations were determined using a NanoDrop 2000c spectrophotometer (ThermoFisher Scientific; Waltham, MA) and were diluted to 10 ng/μL in Type I purified water. PCR primers (Integrated DNA Technologies; Coralville, IA) were prepared with standard desalting and are listed in Table 2-1. A total of 20 ng of DNA was combined with 20 μL of Long-Amp *Taq* Polymerase 5x buffer, 2.5 μL of 10 mM dNTP, 4 μL Long-Amp *Taq* Polymerase(2500U/ml) (all from New England Biolabs), and 2 μL of a single DOP-PCR oligonucleotide primer (100μM) in 100 μL volume in a 300 μL thin-walled PCR tube (ThermoFisher Scientific). Samples labeled "$MnCl_2$" included the addition of 70 μM $MnCl_2$ in the PCR mix. Samples designated "Q5" contained 20 μL Q5® Polymerase 5x Buffer and 2 μL Q5®DNA polymerase (New England Biolabs) instead of the *Taq* polymerase and buffer.

Table 2-1. Primers and sequences examined by DOP-PCR and next-generation sequencing

Primer Name	Sequence	Origin
Rand3	ATGCGACGCCTTAAGGGGGGGGGNNNNNN	Trial-and-error (this study)
BamHIDOP	TGAGGATCCGGGCATGGGGGGGGGNNNNNN	Trial-and-error (this study)
DOPcontrol	CCGACTCGAGNNNNNNATGTGG	Benchmark, Barbaux et al. (2001)
9G_6N	TAGACAATGGGGCATGGGGGGGGGNNNNNN	Trial-and-error (this study)
9NATG	TAGACAATGGGGCATNNNNNNNNNNATG	Trial-and-error, Freedman et al. (2014)
9G_3N	TAGACAATGGGGCATGGGGGGGGGNNN	Trial-and-error (this study)
M1	GAGTGGGGTTGGGTTGGGGGGNNNNNN	NUPACK thermodynamic design (this study)
M2	TATTTCAAATAATGGGGGGGGNNNNNN	NUPACK thermodynamic design (this study)
M4	ATATGTTGAACTTATGGGGGGNNNNNN	NUPACK thermodynamic design (this study)
M5	ATATGTTGAACTTATCCCCCCNNNNNN	NUPACK thermodynamic design (this study)
L1	GAGTGGGGTTGGGTTGGGGGGGGGNNNNNN	NUPACK thermodynamic design (this study)
L2	TATTTCAAATAATGGGGGGGGGGGNNNNNN	NUPACK thermodynamic design (this study)
$MnCl_2$ (L1)	GAGTGGGGTTGGGTTGGGGGGGGGNNNNNN	Primer L1 with 70µM $MnCl_2$ added
Q5 (L1)	GAGTGGGGTTGGGTTGGGGGGGGGNNNNNN	Primer L1 with the Q5 polymerase
Cold5 (L1)	GAGTGGGGTTGGGTTGGGGGGGGGNNNNNN	Primer L1 with 5 low and 30 high stringency cycles

Thermocycling parameters

DOP-PCR reactions were assembled on ice and added to a preheated MasterCycler Gradient Thermocycler (Eppendorf; Hauppauge, NY). The standard DOP-PCR program involved 10 cycles of low stringency amplification involving, first, a 5 min denaturation at 95°C. Then the following 10 cycles were executed: (i) 94°C for 1 min, (ii) 30°C for 2 min, (iii) ramp to 68°C over 3 min, and (iv) amplification at 68°C for 8 min. Following this program, a high stringency program of 25 cycles was implemented. Each cycle consisted of: (i) 94°C for 1 min, (ii) 55°C for 2 min, and (iii) 68°C for 8 min, with 5 s added for each cycle. The DOP-PCR products were held at 68°C for 7 min and then stored at -35°C. The low stringency cycles were designed to encourage random priming, and the high stringency cycles were designed to further amplify those randomly produced DNA fragments. For the samples labeled "Cold5 (L1)," the protocol was adjusted to decrease the number of low stringency cycles to five while increasing high stringency cycles to 30. The total amplification time was approximately 6 h.

Amplified DOP-PCR products were purified using GeneJET PCR purification columns (ThermoFisher Scientific) and eluted in Type I purified water. The purified product concentrations were determined by NanoDrop 2000c. Gel images of the amplified DOP-PCR products were examined on a tris-acetate EDTA agarose gel with ethidium bromide and imaged using a Gel Doc XR System with Image Lab software (BioRad; Hercules, CA). Additional details of the DOP-PCR library generation as well as the next-generation sequencing methods, sequencing data analysis, and genome re-assembly (discussed below) are given in the Supplementary Appendix.

Next-generation sequencing

Next-generation sequencing and sample preparation were performed by the Virginia Bioinformatics Institute in Blacksburg, VA. To start, 1 μg of each DOP-PCR amplified sample (and 1 μg unamplified control) were sheared using an M220 focused ultrasonicator (Covaris; Woburn, MA) and a 50 μL sample containing fragments to 220 bp was analyzed using the Qubit dsDNA HS Assay kit (Life Technologies; Durham, NC) and the Bioanalyzer High Sensitivity kit (Agilent; Santa Clara, CA) to verify size distribution. The sheared DNA fragments were prepared for sequencing using the TruSeq SBS Rapid Duo kit (Illumina; San Diego, CA), followed by one PE Rapid Run cluster kit, one 200 cycle SBS kit, two 50 cycle SBS kits, and a PrepX™ ILM DNA Library Reagent kit (all applied according to manufacturer's directions). Re-amplified samples were loaded to a final concentration of 10 pM in an Illumina® HiSeq 2500 Rapid Run 151SR v3 in two 24 sample lanes with PhiX control v3 diluted to 12.5 pM.

Sequencing data analysis

The "barcode" region (static and degenerate primer regions) of the DOP-PCR amplified DNA fragments remaining from the run and sequencing and required removal computationally with the FASTX Toolkit barcode splitter [40] set to seek the 5' static region of the primer sequence with compensation for two mismatches and one frame-shift. Read trimming was performed using Trimmomatic v0.32 with the following settings: "ILLUMINACLIP 1:40:30 LEADING:3 TRAILING:3 MAXINFO:20:0:2" [41]. Cleaned reads were aligned to the reference genomes (*E. coli*: NC_010473.1, *C. acetobutylicum*: NC_003030.1 (genome) and NC_001988.2 (megaplasmid), and *C. testosteroni*: NZ_AHIL01000001-NZ_AHIL01000063 containing 63 assembled contigs) using the Bowtie2 program [42]. Variants between the

amplified and reference genome were calculated using Samtools mpileup V1.0 [43] and filtered to keep only sequences where the variant calls out-number the reference by 90%. The aligned reads were analyzed with Samtools and Bedtools [44] to merge sets created with primers M1, M2, M4, M5 into a single "Mset" and L1 and L2 Into "Lset", as well as to randomly sample 1M aligned reads from each dataset for analysis.

Genome re-assembly

Reads from each sample that aligned with the reference genome were randomly subsampled into 1M read FASTQ files, and the genome was re-assembled using SPAdes v3.1.0 [45] with a minimum contig size of 200 bp. This was done to eliminate small, potentially erroneous, reads in favor of large genome-aligning sequences. Three samplings and assemblies were performed per sample and the average genomic coverage, mismatch count, N50 (a statistic related to fragment size), and total number of contigs were calculated using QUAST v2.1 [46].

Thermodynamic calculations

The hypothesis driving this research is that DOP-PCR primers for whole genome amplification can be better designed using thermodynamics. Specifically, the goal is to eliminate primers that form hairpin structures or maintain high melting temperatures. To do this, the DOP-PCR primer was considered to have a static sequence on its 5' end and a degenerate sequence of either 6 or 9 random ("N") bases on its 3' end. Several 5' static ends were retrieved from the literature as well as designed and examined with degenerate 3' ends. In preliminary studies, it was also found that a poly guanine ("G," poly-G) static sequence should precede the degenerate sequence (discussed later). For the analysis, the Gibb's free energy of hairpin formation was

calculated between 30°C and 65°C (step-size of 5°C) with monovalent ion concentration of 50 mM and a divalent cation concentration of 2 mM. Calculations were carried out using the stand-alone version of NUPACK software [47] interfaced with MATALB (R2013A). This setup allowed for the Gibb's free energy of hairpin formation and the melting temperature to be calculated across 2,500 5' static "barcode" ends with a 6-9bp poly-guanine or poly-cytosine spacer and degenerate 3' sequence. Static ends with the 1,000 lowest Gibbs free energy of hairpin formation at arbitrary 3' regions were selected and recalculated with 1,000 random 3' sequences. Static ends with the lowest average minimal energy were tested exhaustively against all possible 3' degenerate combinations to locate incompatible sequences. A 6 bp degenerate sequence, for example, was tested in 4^6 (4,096) thermodynamic calculations. It was hypothesized that significant bias occurs in DOP-PCR because a large number of primer combinations hairpin easily, preventing any amplification by those primers. Finally, multiple 5' ends were designed so that every combination of degenerate sequence was included in at least one primer that did not form a hairpin structure. It was hypothesized that this combination of primers would result in minimally biased whole genome amplification by DOP-PCR.

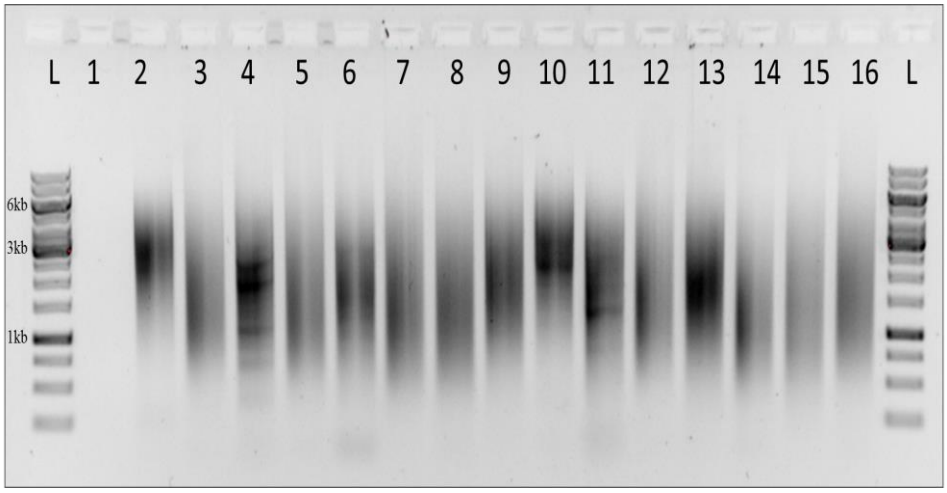


Figure 2-1. DOP-PCR amplification of E. coli NEB 10-beta genomic DNA. Gel image was produced from 3 µL of amplified DNA was run on a 0.8% agarose TAE electrophoresis gel pre-stained with ethidium bromide. Lanes: (L) exACTGene 1 kb ladder, (1) no primer, (2) Rand3, (3) BamHI, (4) DOPcontrol, (5) 9G_6N, (6) 9N_ATG, (7) 9G_3N, (8) M1, (9) M2, (10) M4, (11) M5, (12) L1, (13) L2, (14) $MnCl_2$ (L1) (15) Q5 (L1), (16) Cold5 (L1)

RESULTS

DOP-PCR primer design

Trial-and-error

Numerous oligonucleotide primer designs were tested for DOP-PCR DNA amplification with the goal of achieving a large average fragment size (i.e., gene-sized fragments) and reduced bias over a genome. Different genomes served as the DNA template for DOP-PCR trials: *E. coli* (50.8% GC content), *C. acetobutylicum* (30.9% GC content), and *C. testosteroni* (61.8% GC content). As stated previously, the DOP-PCR approach is important for the creation of genomic DNA libraries, as the DOP-PCR DNA fragments are A-tailed and clone easily with a linearized T-tailed vector [48]. Several primers yielding "successful" DOP-PCR amplification are

presented in Table 2-1. The corresponding DOP-PCR amplification gel image results are shown in Figure 2-1. From the gel image, all primers produce a "smear" of DNA upon DOP-PCR amplification of *E. coli* genomic DNA. This smear contains a heterogeneous mixture of amplified DNA between roughly 1-6 kb in size.

Amplification bias towards certain sequences in a genome can be identified in a gel image as "banding." This is somewhat obvious for the DOPcontrol primer, which was designed in previous research [49]. The DOPcontrol primer was our benchmark, and improvements in reducing genome bias were sought in this research. In the first iterations of primer design, we used a simple "trial-and-error" approach with inferences from previously published accounts. Examples of "unsuccessful" primer DOP-PCR amplification results are shown in Figure 2-2 and contain significant banding, uneven or no amplification, and/or a significant amount of DNA fragments below 1 kb.

The design of the DOPcontrol primer contained a degenerate region followed by a static "ATGTGG" at the 3' end (Table 2-1). This design was amended to form primer 9NATG by keeping the degenerate region (expanded to 9 nucleotides), shortening the static 3' region to "ATG" to attract microbial gene start sites, and changing the 5' static end to be compatible with an expression vector through Gibson cloning methods. Experiments with this primer demonstrated almost exclusive guanine ("G") bias of the degenerate nucleotides when amplifying *E. coli* genomic DNA [48]. This was found by sequencing many amplified and cloned DNA fragments. This bias of "G" was found advantageous due to the decreased melting temperatures (even with mis-matched DNA) found with poly-G sequences. So, in further designs, the poly-G sequence was positioned ahead of the degenerate sequence at the 3' end. This was found to positively influence DOP-PCR amplification and reduce banding in gel

images. The DOP-PCR amplification results of several more primer designs are included in the Supplementary Appendix (Table 7-1). Overall, the trial-and-error approach with a poly-G and degenerate 3' end sequences led to a few successful primer designs (Table 7-1), and the Rand3 primer showed some especially superior characteristics (to be discussed).

Thermodynamics

Thermodynamic calculations were used to expedite the design and further reduce genome amplification bias. Calculations were performed using NUPACK[47] software on a MATALB (R2013A) interface. The objective of the calculations was to predict the Gibb's free energy of hairpin formation (a cutoff of -1 kcal/mol was used), given each possible combination of bases in the degenerate region. The poly-G sequence was conserved in these primers, and the hairpin predictions were also used to design the static 5' end of the primers. From exhaustive calculations, no single primer with a 3' end degenerate region could be located in which hairpin formation at 30-65°C was avoided. When no single primer could be found, the search was expanded to locate sets of primers. With the final design, each combination of the six degenerate nucleotides could be expressed stably in at least one primer. This set of primers is represented as the "Mset," which consists of primers "M1," "M2," "M4," and "M5" (Table 2-1 and Figure 2-1). The "M5" primer contains a poly-C region instead of the poly-G in order to minimize hairpin formation at some degenerate region combinations. An additional set of primers "L1" and "L2" (together form "Lset") were identical to "M1" and "M2" except for an extended poly-G sequence. While several candidate primers were synthesized, tested, and viewed on gels, those in Table 2-1 were applied to all three genomes (*E. coli*, *C. acetobutylicum*, and *C. testosteroni*) and DOP-PCR amplification results were analyzed in detail by next-generation sequencing.

Next-generation sequencing results

DOP-PCR amplification and genome assembly

The DOP-PCR amplified genomic libraries were sequenced, and results are discussed in the next several sections. Summaries of data are collected in Tables 2 and 3 as well as in Tables S2 and S3 in the Supplementary Appendix. The percentage of reads that were aligned to the genome, the coverage of the reference genome (at 1x and 10x), and the N50 statistic (a measure of the largest DNA fragment size for which 50% of the assemble fragment lengths are shorter) for the re-assembled reference genomes are given in Table 2-2 for every primer and genome combination. The GC content of the libraries, the Kolmogorov-Smirnov (K-S) test statistic (a measure of DOP-PCR amplification bias), and the amplification error rate are given in Table 2-3. Additional details regarding number of total reads, the number of aligned reads, and the amount of amplification of DNA during DOP-PCR (determined by both NanoDrop and Qubit) are given in Table 7-2 for every primer and genome combination. Finally, information on the number of full genes, partial genes, and the number of contigs present in genome re-assemblies is given in Table 7-3 of the Supplementary Appendix. Following DOP-PCR amplification and column purification (to recover DNA fragments of 100 bp to 10 kb and remove primers, residual DNA/RNA, proteins, etc.), DNA concentrations were determined by NanoDrop™ 2000c using absorbance at 260 nm. Based on an initial genomic template concentration of 200 pg/μL, the final DNA libraries were amplified 20-600x (Table 7-2). Amplification was highest for primers utilizing the 6 bp degenerate region on the 3' end. Amplification yields based on Qubit® measurements were 10-100 fold higher, approaching 10,000x amplification for *C. testosteroni*,

which contained 3 pg/μL DNA in each DOP-PCR amplification, demonstrating the sensitivity of this technique.

Genome assembly was performed using aligned sequencing reads. The assembly cut-off was 200bp, slightly larger than a single sequence read. If a read originated from a DOP amplified fragment, it should ideally be 500-6000bp in length with a primer 5' binding site on each end. Single, isolated reads in the genome could not have arisen from such an amplified fragment and were excluded from assembly coverage calculations. The results of QUAST v2.1 indicated similar genomic coverage trends to the Bowtie2 alignments. Results of genome assembly specifically highlighted (i) advantages of the longer degenerate region, (ii) advantages of grouping thermodynamically designed primer sets together (Mset), (iii) the weakness of a static region located at the 3' end (e.g., DOPcontrol and 9NATG), and (iv) a slight advantage of the smaller poly-G region (M primers vs. L primers).

Table 2-2. Summary of aligned reads, genome coverage, and N50 for DNA libraries produced by DOP-PCR.

	***E. coli* NEB 10-beta (50.8% GC content)**			***C. acetobutylicum* ATCC 824 (30.9% GC content)**			***C. testosteroni* ATCC 11996 (61.8% GC content)**		
Primer Name	Aligned Reads (%)	Reference Genome Coverage (%) at 1x (10x)	N50[3]	Aligned Reads (%)	Reference Genome Coverage (%) at 1x (10x)	N50[3]	Aligned Reads (%)	Reference Genome Coverage (%) at 1x (10x)	N50[3]
None	98.2	99 (97)	12,347	98.3	100 (99)	18,477	96.2	100 (95)	8,964
Rand3	56.1	97 (58)	2,993	60.4	82 (38)	4,180	83.5	95 (54)	3,098
BamHI	61.1	92 (47)	2,022	29.4	74 (16)	674	26.8	98 (55)	3,752
DOPcontrol	71.8	62 (20)	1,489	97.5	49 (17)	1,655	27.6	72 (21)	1,269
9G_6N	66.0	94 (48)	2,266	92.9	67 (28)	2,162	46.3	94 (48)	2,903
9NATG	44.9	84 (31)	2,078	82.9	51 (15)	1,817	3.8	N/A[1]	N/A[1]
9G_3N	82.8	75 (32)	1,744	79.1	30 (12)	1,578	62.1	81 (32)	1,421
M1	51.4	90 (44)	1,866	76.8	64 (25)	1,528	30.8	91 (38)	1,768
M2	37.9	95 (44)	2,854	43.2	83 (33)	1,692	40.6	93 (43)	2,951
M4	25.3	96 (46)	2,377	79.7	78 (31)	2,225	24.0	94 (45)	1,532
M5	2.1	N/A[1]	N/A[1]	4.0	N/A[1]	N/A[1]	32.2	88 (39)	2,747
L1	62.3	92 (38)	2,024	81.7	88 (35)	1,795	86.1	86 (33)	1,219
L2	0.5	N/A[1]	N/A[1]	93.9	71 (30)	2,147	16.2	83 (31)	1,574
$MnCl_2$ (L1)	57.2	90 (33)	1,849	19.8	39 (12)	796	38.7	84 (29)	1,122
Q5 (L1)	97.1	79 (33)	1,172	96.9	57 (24)	2,096	94.7	34 (14)	1,373
Cold5 (L1)	68.1	89 (39)	2,321	73.8	59 (24)	1,824	53.1	92 (36)	1,451
Mset	-	97 (51)	4,710	-	84 (36)	2,731	-	96 (51)	6,485
Lset	-	92 (38)	1,980	-	84 (34)	2,474	-	87 (34)	1,924

[1] Amplifications producing less than 1M aligned reads were excluded from analysis.
[2] Kolmogorov-Smirnov (K-S) test statistic comparing the read distribution of DOP-PCR amplified samples to the unamplified and sheared control.
[3] The N50 is defined as the contig length where 50% of total assembled contig lengths are the specified length or longer

Table 2-3. Summary of GC content, bias and amplification errors for DNA libraries produced by DOP-PCR.

	***E. coli* NEB 10-beta (50.8% GC content)**			***C. acetobutylicum* ATCC 824 (30.9% GC content)**			***C. testosteroni* ATCC 11996 (61.8% GC content)**		
Primer Name	Library GC% Relative to Genome[2]	K-S Statistic[3]	Mismatches Per 100 kb	Library GC% Relative to Genome[2]	K-S Statistic[3]	Mismatches Per 100 kb	Library GC% Relative to Genome[2]	K-S Statistic[3]	Mismatches Per 100 kb
None	0.06	0	3.9	-0.4	0	8.4	0.1	0	5.8
Rand3	2.1	0.447	48.8	17	0.658	72.8	-2.0	0.533	51.3
BamHI	1.7	0.519	56.5	37.7	0.786	87.8	-2.4	0.546	46.5
DOPcontrol	-0.5	0.764	81.1	4.1	0.802	87.9	-5.0	0.802	117.4
9G_6N	0.6	0.515	60.5	21.2	0.712	81.8	-3.0	0.598	49.9
9NATG	1.1	0.657	71.2	3.1	0.847	68.3	N/A[1]	N/A[1]	N/A[1]
9G_3N	-0.1	0.645	72.1	15.4	0.834	106.7	-4.5	0.710	90.1
M1	-0.4	0.557	60.9	13.2	0.739	90.6	-8.0	0.682	68.3
M2	0.4	0.574	62.5	14.1	0.686	87.8	-3.6	0.628	67.9
M4	-1	0.559	83.3	9.3	0.692	72.2	-4.9	0.601	96.6
M5	N/A[1]	N/A[1]	N/A[1]	N/A[1]	N/A[1]	N/A[1]	-5.8	0.714	80.5
L1	0.6	0.610	79.8	18	0.670	86.3	-5.5	0.650	90.6
L2	N/A[1]	N/A[1]	N/A[1]	9.9	0.699	83.1	-5.0	0.742	99.1
$MnCl_2$ (L1)	0.6	0.657	178.6	31.9	0.827	253.4	-6.6	0.729	221.3
Q5 (L1)	0.7	0.649	43.4	14.5	0.748	48.7	-11.1	0.904	95.7
Cold5 (L1)	0	0.599	56.0	13.8	0.745	96.0	-5.1	0.652	98.8
Mset	-	0.504	48.7	-	0.664	69.7	-	0.607	47.9
Lset	-	0.609	78.9	-	0.672	81.4	-	0.659	84.5

[1] Amplifications producing less than 1M aligned reads were excluded from analysis.
[2] This is the percent GC content of the DNA library above (positive value) or below (negative value) the published value for the reference genome.
[3] Kolmogorov-Smirnov (K-S) test statistic comparing the read distribution of DOP-PCR amplified samples to the unamplified and sheared control.

Genome alignment

All cleaned single-end reads were aligned to the reference genomes. The raw counts of the total number of reads and aligned reads for each primer with each genome are contained in Table 7-2. From these data, the percentage of aligned reads was calculated and is reported in Table 2-2. The effect of primer composition and genome GC content both had an effect on the overall alignment of reads. Unamplified and sheared genomic DNA (the control; labeled as "None" in all Tables) aligned with >96% success for all genomes; however, the DOP-PCR amplified DNA contained many non-reference reads. The reads that failed to align were either sequences from contaminating DNA, amplification artifacts, or reads which aligned to two separate, unique regions of the reference. Upon closer inspection, this phenomenon occurred in regions of 4-8 bp sequence similarity between both stretches of the genome, leading to the conclusion that an amplified genome fragment self-primed a distant location on the genome (data not shown). The amplified libraries from *E. coli* showed poor alignment (<50%) with the 9NATG and M2, M4, M5, and L2 primers. The Q5 polymerase, which has proofreading function, showed very high alignment across all three genomes where 95-98% of reads aligned to a single portion of the reference genome. Adding $MnCl_2$ to the amplification mix tended to reduce alignment ratios, as did reducing the number of non-stringent annealing cycles (Cold5, L1). The L2 primer showed low amplification alignment for all genomes except for the low GC content *C. acetobutylicum* reference. The *C. acetobutylicum* genome amplified with >75% of library fragments aligning to the genome for the DOPcontrol, 9G_6N, 9G_3N, M1-5, and L2 primers, which were higher rates than for the other genomes. *C. testosteroni* libraries using the Rand3 and L1 primers produced the best alignments. Amplification using primers with a poly-G region length of 9 produced greater alignment than the poly-G region length of 6. There was no

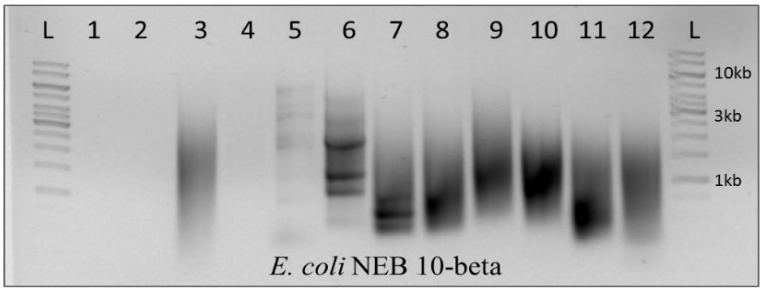
clear alignment advantage in using 3 or 6 degenerate nucleotides on the 3' end of the oligonucleotide primers, and comparing the 9NATG and 9G_3N primers indicated that internal degenerate nucleotides resulted in lower alignment than terminal ones. For comparison, the benchmark DOPcontrol primer showed significant bias to the GC content of the genome with 71.8% alignment for *E. coli* (50.8% GC content), 97.5% alignment for *C. acetobutylicum* (low GC), and only 27.6% alignment for *C. testosteroni* (high GC). However, as will be discussed, poor alignment does not necessarily correlate with poor genomic coverage. Sequencing runs that failed to produce at least 1M aligned reads were not included in genome assembly or further analyses. It is likely that these samples were hindered by excessive primer-dimer formation.

Genome coverage

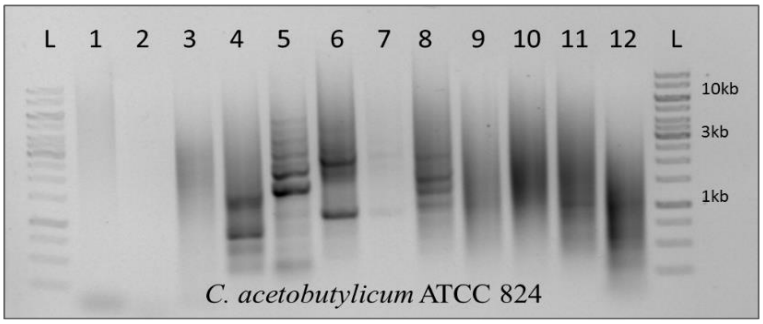
Genomic coverage was calculated for each primer with each genome. Genomic coverage was analyzed by aligning the genome assembly to the NCBI General Feature File (gff) for that genome. Since these files only exist for *E. coli* and *C. acetobutylicum*, manual curation was performed for *C. testosteroni* from the NCBI GeneBank file. Coverage results correlate with the number of genes represented. The gene metric is important to a genomic library due to the desire to test for expression of entire genes and coding sequences. Alignment files were sampled randomly for 1M reads and the coverage breakdown is given in Table 2-2. Unamplified and sheared genomic DNA (the control) showed the most uniform coverage, where 1x coverage refers to bases in the genome represented at least once by the sequencing data and 10x coverage refers to how many bases are represented by at least 10 reads. Except for the genomic DNA, coverage fell rapidly between 1x and 10x. The *E. coli* genome resulted in higher 1x and 10x coverage for virtually all primers than the other genomes, suggesting challenges exist in

uniformly amplifying genomes with low and high GC contents. Of particular interest is that several of the trial-and-error and thermodynamically designed primers showed a percent coverage significantly higher (up to 100% higher in some cases) than the benchmark DOPcontrol primer. This suggests that the methods developed for screening primers due to the presence of banding in the gel image and the thermodynamic design was effective in reducing genome bias compared to the previously published benchmark DOP-PCR primer. Combining the "M" and "L" primers into a single set (Mset and Lset) produced mixtures with improved properties compared to the individual primers. Also, a primer design with a longer degenerate region consistently out-performed a primer with a smaller degenerate region for all genomes. The extended poly-G region led to a slight increase in coverage, and while the Q5 polymerase led to superior alignment results, the coverage results were largely inferior. In addition, the Rand3 primer performed particularly well for all genomes.

A



B



C

Lane	Primer Sequence
1	TATATATATATATATNNNNNNATG
2	CGCTATCGCTATAGCNNNNNNATG
3	CGCCGCCGCCGCCGCNNNNNNATG
4	CGCTATCGCTATAGCGGGGGGATG
5	TAGACAATGGGGCAGGGGGGGGGGGG
6	CGCTATCGCTATAGCGGGGGGNNN
7	TAGACAATGGGGCATGGGGGGGGGATG
8	TAGACAATGGGGCATGGGGGGGGGNN
9	TAGACAATGGGGCATGGGGGGGGGNNN
10	TAGACAATGGGGCATGGGGGGGGGNNNNNN
11	TAGACAATGGGGCATGGGGGGNNN
12	TAGACAATGGGGCATGGGGGGNNNNNN

Figure 2-2. Examples of "unsuccessful" DOP-PCR amplification marked by absence of amplification, "banding," uneven amplification, and abundance of DNA fragment of size < 1 kb. (A) *E. coli* amplification, (B) *C. acetobutylicum* amplification, (C) primers used in lanes 1-12. The Exactgene 1 kb ladder is shown in lane L.

Amplified DNA fragment size

Assembling a uniform number of aligned reads into larger contigs was performed based on stretches of DNA fragment overlap. This allowed calculation of the average size of each covered region of the genome. The number of independent contigs of each assembled genome for DNA libraries produced by each primer is given in Table 7-3 along with the numbers of full and partial genes present in each genome re-assembly. The size of DNA fragment at which 50% of the genome is represented by fragments of equal or greater length is referred to as the N50 [46], and this statistic is given in Table 2-2.

In addition, gel images (Figs. 1 and 2) also offer a visualization of the upper and lower bounds of DNA fragment size as well as the overall distribution. The number of contigs

generated for 1M reads for each assembly was performed using SPAdes v3.1.0 [45]. Results showed that the unamplified DNA control (labeled "None") could be assembled into the fewest contiguous elements with a size range of 8-18 kb. The N50 for the unamplified *E. coli* genomic DNA exceeded 12 kb, and this value exceeded 18 kb and 8 kb for the *C. acetobutylicum* and *C. testosteroni* genomes, respectively. The genomes assembled from DOP-PCR amplified DNA yielded smaller N50 values and a larger number of total contigs. Interestingly, the Rand3 primer was a top-performer for all 3 genomes in terms of (i) high N50 value (large stretches of genome coverage), (ii) low number of contigs, and (iii) high number of full genes represented in the DNA library. Consistent with the gel image (Figure 2-1), the Rand3 primer amplified products yielding large fragment sizes with high concentrations in the 3-6 kb size range. The *C. testosteroni* genome amplified products were largest under the BamHI primer, which clearly showed preference for high GC content of the genome. The Mset mixture of primers gave higher N50 values, relative to the individual "M" primers. In particular, the Mset offered improvements in N50 of 154%, 79%, and 267% relative to the M1 primer for the *E. coli*, *C. acetobutylicum*, and *C. testosteroni* genomes, respectively. Additionally, the Mset resulted in fewer contigs and more full genes (Table 7-3). The Mset out-performed Rand3 in the neutral GC content genome of *E. coli*, and especially with the high GC content genome of *C. testosteroni*; however, the Rand3 primer was better with the low GC content genome of *C. acetobutylicum*, further suggesting the GC content of the reference genome plays a large role in determining DOP-PCR amplification effectiveness with different primers. The benefit of including a 6 residue poly-G region (as opposed to 9 residues) was seen for the *E. coli* and *C. testosteroni* genomes (but not for *C. acetobutylicum*) in terms of N50 and coverage. This may be due to the incompatibility of long poly-G regions to stabilize interactions with a low GC content genome, forcing more unfavorable

G:A and G:T interactions that discourage primer annealing and amplification. Clear advantages were also observed by using 6 degenerate nucleotides (as opposed to 3 or 9) for otherwise identical primers. Also important to this research is that almost all primers listed in Table1 out-performed the DOPcontrol primer in terms of both genome coverage and fragment size.

Library GC content

The GC content of the reference genome affects that of the DOP-PCR amplified DNA library; however, the GC content of the library was found to be disproportionate to that of the reference genome. For the DNA library produced from *E. coli* genomic DNA, the GC content of the sequenced reads were within 1% of the reference genome GC content, which is to be expected. Most variance was seen from the Rand3 primer. However, the GC content variation in DOP-PCR amplified DNA increased for both the *C. acetobutylicum* and *C. testosteroni* genomes. The amplified DNA from the low GC *C. acetobutylicum* genome was consistently higher than the GC content of the genome (30.9%). Here, the BamHI and the $MnCl_2$ (L1) amplifications led to the most significant deviations (in excess of 30% difference from the genome). For the DNA amplified from the high GC *C. testosteroni* genome, the GC content was consistently below that of the genome (61.8%). However, only one primer run (Q5 (L1)) showed more than a 10% difference from the genome.

Amplification bias

The DOP-PCR amplification failed to produce completely unbiased DNA libraries, in agreement with previous bias results [35], but improvements were seen using the thermodynamic primer design. An unbiased DNA library has its DNA fragments distributed uniformly

throughout the genome, resulting in a normal probability distribution. The Kolmogorov-Smirnov (K-S) test statistic was calculated based on start sites of aligned DNA fragments for all primers with all genomes. The K-S statistic describes the difference in one-dimensional probability distributions. Here, this allowed for the calculation of DOP-PCR amplification bias, as the DNA libraries were compared to the unamplified and sheared control, which was found to follow a normal distribution. These results are also presented in Table 2-3, where K-S statistic values closer to zero signal less bias in the DNA library. The distribution plots are also presented in Figs. S2-S4 for all primers and genomes. Overall, none of the primers showed equivalent bias to the control (given α=0.05). However, almost all primers designed in this research and listed in Table 2-3 out-performed the benchmark DOPcontrol primer. In particular, the Rand3 primer showed consistently the lowest K-S statistic value across all three genomes and outperformed the thermodynamically designed Mset by about 10%. In addition, the Mset of primers, on the other hand, out-performed the individual M1-5 primers for all three genomes and had a K-S statistic that out-performed that of the DOPcontrol by as much as 35%. The influence of genome GC content was also seen, as the K-S statistic values for *E. coli* (50.8% GC content) were lower than for either the low GC *C. acetobutylicum* or high GC *C. testosteroni* genomes.

Amplification errors

The utility of DOP-PCR as a means of SNP detection from small quantities of DNA has been examined [50]. While the previous research suggests DOP-PCR typically results in a 0.7% error rate (700 DNA mismatches per 100 kb), the research presented here with thermodynamically designed primers appears much more accurate, with most primers showing less than 100 mismatches per 100 kb, when results were calculated from re-assembled genomes.

Mismatches (per 100 kb) were lowest for the unamplified and sheared DNA (less than 10 per 100 kb). The proofreading Q5 polymerase also reduced errors when amplifying the *E. coli* and *C. acetobutylicum* genomes but not with the high GC *C. testosteroni* genome. In addition, the DOPcontrol benchmark primer led to one of the highest error rates, while the Mset and Rand3 primers produced among the lowest mismatch rates. The differences between the two are 67%, 26%, and 145% for the *E. coli*, *C. acetobutylicum*, and *C. testosteroni* genomes, respectively, and a GC content bias is apparent. Error rate could be further controlled by the addition of 70 μM $MnCl_2$ to the DOP-PCR reaction mixture in run "$MnCl_2$ (L1)". This enabled the purposeful induction of errors to generate "error-prone" DNA libraries, which may be screened for enhanced capabilities [51]. In comparison to the L1 primer without manganese, the amplified product had increased error rates of 123%, 194%, and 144% for the three genomes, respectively.

DISCUSSION

Genomic library construction by DOP-PCR

Primer design for DOP-PCR method for DNA amplification was examined in this research for the purpose of generating genomic DNA libraries. We have described methods by which DOP-PCR generated DNA fragments can be cloned into expression vectors with very high efficiency for effective library screening [48]. This offers an alternative approach to traditional methods of genomic library construction that involves physical shearing and end-polishing of DNA or biased digestion before cloning. Both of these methods resulted in poor cloning efficiency of DNA fragments in our laboratory experiments prior to attempting the DOP-PCR alternative. In addition to their ease of construction, DOP-PCR generated DNA libraries avoid some types of biases of traditional libraries (while having their own due to selective

amplification). Fragmentation of whole genomes by sonication, enzymatic digestion, or nebulization yields tighter control of fragment size distribution but fails to produce truly random stand breaks, as the composition of the genome can influence the location of strand breakage under common shear conditions, predominantly affecting GC rich regions more than AT regions [19]. DNA fragmentation by enzymatic digestion occurs at defined sites (leading to loss of randomness) but the digestion frequency can be controlled by the amount of enzyme added. Further, shearing or sonication of the DNA involves the additional end-polishing step to blunt single-stranded overhangs, addition of a phosphate group for ligation, or overhangs for TA cloning. The DOP-PCR approach provides a one-step protocol to produce DNA library fragments that are ready be cloned efficiently into a T-tailed vector. However, cloning of DOP DNA enzymatically digested with a BamHI (using primer BamHI, Table1) also led to inefficiency and low yield in our experience.

Another important advantage of DOP-PCR amplified libraries is the amount reference DNA required. Generation of genomic libraries through traditional approaches can require multiple micrograms of DNA before shearing. Harvesting enough genomic DNA can be challenging for the vast majority of non-laboratory strains in nature and for situations (such as forensics) where the DNA is non-replicating. In our experience, DOP-PCR has achieved over 1,000x amplification of a genome, creating 3 μg of amplified DNA fragments for cloning from a starting amount of 3 ng of DNA.

Aspects of primer design

DOP-PCR amplification related biases were investigated in this research in order to design new sets of primers. Throughout this research, the new designs were compared to a

published benchmark DOP-PCR primer referred to as "DOPcontrol" [49] (Table 2-1). By testing multiple primer designs, amplification conditions, thermostable polymerases, and genomes of varying GC content, we aimed to optimize the DNA fragment size obtained from DOP-PCR while minimizing the amplification bias. Through re-designing the DOP-PCR primer, we have shifted the location of the degenerate nucleotide region from the interior of the primer to the 3' end, allowing for more randomized amplification start sites (compared to the ATGTTG 3' sequence of DOPcontrol). The addition of a thermostable region of a static poly-G or poly-C region is believed to have stabilized the degenerate region, possibly explaining success using 6 (or 9) degenerate nucleotides, where this has been unsuccessful previously [22]. As the degenerate nucleotides have the potential to take on any theoretical sequence, care must be taken with primer designs that will self-anneal between the degenerate and static regions. In fact, all static regions will anneal with some combination of the degenerate region. Through use of the NUPACK software and MATLAB, primer designs were developed that work in tandem, allowing practically all combinations of the degenerate sequences to exist unbound somewhere in the primer mixture. However, some degenerate combinations will form hairpins more easily than others, meaning they are not all available for DNA amplification in equal abundance. This, of course, can lead to bias. With issues such as this, it is apparent that much room for further optimization of DOP-PCR primer sets remains. This research represents a significant step forward in primer design, but future improvements are anticipated.

What primers should be used for DOP-PCR DNA amplification?

With the information presented here, how would one most effectively create a DNA library using DOP-PCR? It appears that primers are sensitive to the GC content of the reference

genome, so it would make sense that primers should be designed based on this criterion. However, it is difficult to know the GC content of, for example, a soil sample metagenome without initial testing. Thus, one goal of this research was to design primer sets that would be effective across the broad range of potential GC contents. When faced with the question of which primers to use after the next-generation sequencing results were processed, we chose to use a combination of the Mset of primers with the Rand3 primer in several parallel DOP-PCR amplification runs to generate a comprehensive DNA library for cloning. Our results confirm suspicions of previous groups [22] that running several DOP-PCR reactions with separate primers and pooling decreases library coverage bias only when the primers are designed with thermodynamic constraints. When using the combination of the Mset and Rand3, calculations showed an increase in genome coverage, more favorable N50 values, and K-S statistics with less (but still significant) amplification bias compared to the other choices of primers. In addition, this method of thermodynamic primer design and pooling out-performed the benchmark DOPcontrol primer considerably in every DNA library assessment analyzed. This method of DOP-PCR genomic DNA library construction and cloning has been reproducible and consistent in our experience, it can be completed by researchers (non-specialists) with working knowledge of PCR and TA cloning, and it incurs minimal cost. This represents a step forward for DNA library construction (especially from metagenomes) and cloning and has the potential to bring DNA library screening to several new laboratories.

ACKNOWLEDGEMENTS

Research funding is acknowledged from the USDA CSREES AFRI (Award: 2010-65504-20346), the Jeffress Trust Awards Program in Interdisciplinary Research, and the Institute

of Critical Technologies and Applied Science at Virginia Tech. Computational resources were provided by the Virginia Tech Advanced Research Computing Cluster. We also acknowledge contributions form Nina Wilson, Parker W. Lee, and Jessica Symons and. From the Virginia Bioinformatics Institute, we acknowledge Jennifer Jenrette, Robert Settlage, and Saikumar Karyala.

REFERENCES

[1] Rommens, J., Iannuzzi, M., Kerem, B., Drumm, M., *et al.*, Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989, *245*, 1059-1065.

[2] Clarke, L., Carbon, J., A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire E. coli genome. *Cell* 1976, *9*, 91-99.

[3] Kennedy, J., Marchesi, J. R., Dobson, A. D., Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb Cell Fact* 2008, *7*, 27.

[4] James, P., Halladay, J., Craig, E. A., Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* 1996, *144*, 1425-1436.

[5] Jauert, P. A., Jensen, L. E., Kirkpatrick, D. T., A novel yeast genomic DNA library on a geneticin-resistance vector. *Yeast* 2005, *22*, 653-657.

[6] Nicolaou, S. A., Gaida, S. M., Papoutsakis, E. T., Coexisting/Coexpressing Genomic Libraries (CoGeL) identify interactions among distantly located genetic loci for developing complex microbial phenotypes. *Nucleic Acids Research* 2011, *39*, e152-e152.

[7] Woodruff, L. B., Pandhal, J., Ow, S. Y., Karimpour-Fard, A., *et al.*, Genome-scale identification and characterization of ethanol tolerance genes in Escherichia coli. *Metab Eng* 2013, *15*, 124-133.

[8] Borden, J. R., Papoutsakis, E. T., Dynamics of genomic-library enrichment and identification of solvent tolerance genes for Clostridium acetobutylicum. *Applied and environmental microbiology* 2007, *73*, 3061-3068.

[9] Glebes, T. Y., Sandoval, N. R., Reeder, P. J., Schilling, K. D., *et al.*, Genome-wide mapping of furfural tolerance genes in Escherichia coli. *PloS one* 2014, *9*, e87540.

[10] Ceragioli, M., Mols, M., Moezelaar, R., Ghelardi, E., *et al.*, Comparative Transcriptomic and Phenotypic Analysis of the Responses of Bacillus cereus to Various Disinfectant Treatments. *Applied and Environmental Microbiology* 2010, *76*, 3352-3360.

[11] Eyers, L., George, I., Schuler, L., Stenuit, B., *et al.*, Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol* 2004, *66*, 123-130.

[12] Mittler, R., Oxidative stress, antioxidants and stress tolerance. *Trends in Plant Science* 2002, *7*, 405-410.

[13] Alsaker, K. V., Spitzer, T. R., Papoutsakis, E. T., Transcriptional Analysis of spo0A Overexpression in Clostridium acetobutylicum and Its Effect on the Cell's Response to Butanol Stress. *Journal of Bacteriology* 2004, *186*, 1959-1971.

[14] Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E., House, C. H., Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proceedings of the National Academy of Sciences* 2008, *105*, 10583-10588.

[15] Hawkins, T. L., Detter, J. C., Richardson, P. M., Whole genome amplification — applications and advances. *Current Opinion in Biotechnology* 2002, *13*, 65-67.

[16] Peterson, S. N., Snesrud, E., Schork, N. J., Bretz, W. A., Dental caries pathogenicity: a genomic and metagenomic perspective. *International dental journal* 2011, *61 Suppl 1*, 11-22.

[17] Hanson, E. K., Ballantyne, J., Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. *Analytical Biochemistry* 2005, *346*, 246-257.

[18] Russell, J. R., Huang, J., Anand, P., Kucera, K., *et al.*, Biodegradation of Polyester Polyurethane by Endophytic Fungi. *Applied and Environmental Microbiology* 2011, *77*, 6076-6084.

[19] Poptsova, M. S., Il'icheva, I. A., Nechipurenko, D. Y., Panchenko, L. A., *et al.*, Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* 2014, *4*.

[20] Sambrook, J., Russell, D., Irwin, N., Molecular cloning: a laboratory manual-/Joseph Sambrook, David W. Russell;[assoc. ed.: Nina Irwin, Kaaren A. Janssen]. 2001.

[21] Peng, W., Takabayashi, H., Ikawa, K., Whole genome amplification from single cells in preimplantation genetic diagnosis and prenatal diagnosis. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 2007, *131*, 13-20.

[22] Martino, A. J., Rhodes, M. E., Biddle, J. F., Brandt, L. D., *et al.*, Novel degenerate PCR method for whole-genome amplification applied to Peru Margin (ODP Leg 201) subsurface samples. *Frontiers in Microbiology* 2012, *3*.

[23] Biddle, J. F., White, J. R., Teske, A. P., House, C. H., Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. *ISME J* 2011, *5*, 1038-1047.

[24] Rodrigue, S., Malmstrom, R. R., Berlin, A. M., Birren, B. W., *et al.*, Whole Genome Amplification and *De novo* Assembly of Single Bacterial Cells. *PLoS ONE* 2009, *4*, e6864.

[25] Lasken, R. S., Genomic sequencing of uncultured microorganisms from single cells. *Nat Rev Micro* 2012, *10*, 631-640.

[26] Telenius, H. k., Carter, N. P., Bebb, C. E., Nordenskjo¨ld, M., *et al.*, Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* 1992, *13*, 718-725.

[27] Zheng, Y.-m., Wang, N., Li, L., Jin, F., Whole genome amplification in preimplantation genetic diagnosis. *Journal of Zhejiang University SCIENCE B* 2011, *12*, 1-11.

[28] Kittler, R., Stoneking, M., Kayser, M., A Whole Genome Amplification Method to Generate Long Fragments from Low Quantities of Genomic DNA. *Analytical Biochemistry* 2002, *300*, 237-244.

[29] Pan, X., Chen, W., Jia, X., Dong, P., Liang, X., Evaluation of a Whole Genome Amplification Method Based on Improved Ligation-Mediated PCR, in: Zhang, T.-C., Ouyang, P., Kaplan, S., Skarnes, B. (Eds.), *Proceedings of the 2012 International Conference on Applied Biotechnology (ICAB 2012)*, Springer Berlin Heidelberg 2014, pp. 1333-1343.

[30] Kiss, C., Kost-Alimova, M., Klein, G., Szekely, L., Optimisation of the degenerate oligonucleotide primed PCR (DOP-PCR) for capillary thermocycler. *Biomolecular engineering* 2002, *19*, 31-34.

[31] Bonnette, M. D., Pavlova, V. R., Rodier, D. N., Thompson, L. P., *et al.*, dcDegenerate Oligonucleotide Primed-PCR for Multilocus, Genome-wide Analysis From Limited Quantities of DNA. *Diagnostic Molecular Pathology* 2009, *18*, 165-175.

[32] Zhang, L., Cui, X., Schmitt, K., Hubert, R., *et al.*, Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences* 1992, *89*, 5847-5851.

[33] Cheung, V. G., Nelson, S. F., Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proceedings of the National Academy of Sciences* 1996, *93*, 14676-14679.

[34] Dean, F. B., Hosono, S., Fang, L., Wu, X., *et al.*, Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* 2002, *99*, 5261-5266.

[35] Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., *et al.*, Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics* 2006, *7*, 216.

[36] Tsai, I. J., Hunt, M., Holroyd, N., Huckvale, T., *et al.*, Summarizing Specific Profiles in Illumina Sequencing from Whole-Genome Amplified DNA. *DNA Research* 2014, *21*, 243-254.

[37] Kang, M.-J., Yu, H., Kim, S.-K., Park, S.-R., Yang, I., Quantification of trace-level DNA by real-time whole genome amplification. *PLoS ONE* 2011, *6*, e28661.

[38] Harris, L., Blank, L., Desai, R., Welker, N., Papoutsakis, E., Fermentation characterization and flux analysis of recombinant strains of Clostridium acetobutylicum with an inactivated solR gene. *Journal of Industrial Microbiology and Biotechnology* 2001, *27*, 322-328.

[39] Harris, L. M., Welker, N. E., Papoutsakis, E. T., Northern, Morphological, and Fermentation Analysis of spo0A Inactivation and Overexpression in Clostridium acetobutylicum ATCC 824. *J. Bacteriol.* 2002, *184*, 3586-3597.

[40] Gordon, A., Hannon, G., Fastx-toolkit: FASTQ/A short-reads pre-processing tools. 2010.

[41] Bolger, A. M., Lohse, M., Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, btu170.

[42] Langmead, B., Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012, *9*, 357-359.

[43] Li, H., Handsaker, B., Wysoker, A., Fennell, T., *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, *25*, 2078-2079.

[44] Quinlan, A. R., Hall, I. M., BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, *26*, 841-842.

[45] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 2012, *19*, 455-477.

[46] Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013, *29*, 1072-1075.

[47] Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., *et al.*, NUPACK: analysis and design of nucleic acid systems. *Journal of computational chemistry* 2011, *32*, 170-173.

[48] Freedman, B. G., Zu, T. N. K., Wallace, R., Senger, R. S., Raman spectroscopy characterizes 1-butanol tolerant E. coli phenotypes resulting from genomic DNA library enrichment. *Under Review* 2014.

[49] Barbaux, S., Poirier, O., Cambien, F., Use of degenerate oligonucleotide primed PCR (DOP-PCR) for the genotyping of low-concentration DNA samples. *J Mol Med* 2001, *79*, 329-332.

[50] Grant, S. F. A., Steinlicht, S., Nentwich, U., Kern, R., *et al.*, SNP genotyping on a genome-wide amplified DOP-PCR template. *Nucleic Acids Research* 2002, *30*, e125.

[51] Cirino, P. C., Mayer, K. M., Umeno, D., Generating mutant libraries using error-prone PCR. *Methods Mol Biol* 2003, *231*, 3-9.

Chapter 3

Raman spectroscopy characterizes 1-butanol tolerant *E. coli* phenotypes resulting from genomic DNA library enrichment

ABSTRACT

Raman spectroscopy was used to analyze the phenotypes of *E. coli* cells with improved 1-butanol tolerance resulting from genomic DNA library enrichment. DNA libraries were created using degenerate oligonucleotide primed PCR (DOP-PCR) from the genomic DNA of: (i) *E. coli* NEB 10-beta, (ii) *Clostridium acetobutylicum* ATCC 824, and (iii) the metagenome of an environmental soil sample from a stream bank on the Virginia Tech campus. DNA library generation by DOP-PCR presents significant advantages including: (i) high efficiency and repeatable cloning, (ii) ability to use un-culturable organisms or small amounts of template DNA, and (iii) no DNA shearing or end-polishing is necessary. DNA libraries were expressed in *E. coli* and enriched in the presence of 0.9% and 1.1% (v/v) 1-butanol. Strains were isolated with significantly improved 1-butanol tolerance and some had maximum growth rates that were inhibited 50% less than a control strain when grown in the presence of 1-butanol. Nine strains harboring different enriched DNA fragments were analyzed by Raman spectroscopy to reveal vastly different phenotypes can confer increased 1-butanol tolerance. All DNA fragments conferred increased membrane rigidity by increasing the fraction of unsaturated fatty acids at the expense of cyclopropane fatty acids. Strains differed in their responses of accumulating saturated fatty acids as well as total fatty acids relative to cell protein content. This is the first instance of using Raman spectroscopy to scan phenotypes of cells surviving a serial enrichment assay.

Raman spectroscopy data can be obtained in minutes and with minimal sample preparation, potentially enabling high-throughput phenotype screening.

INTRODUCTION

1-Butanol fermentation

Production of the potential biofuel 1-butanol in microbial fermentation is limited to 13-20 g/L (~2% v/v) after which cell metabolism ceases in wild-type *Clostridium acetobutylicum* due to toxicity [1]. This leads to low product titers in fermentation broth and accounts for more difficult separations and increased production costs [2]. *C. acetobutylicum* ATCC 824, the model of acetone/1-butanol/ethanol (ABE) fermentation, and *E. coli*, a common host for engineered metabolic pathways, have been the subjects of metabolic engineering research aimed at increasing alcohol production and tolerance [3-5]. While production strategies such as 2-phase fermentation [6] and 1-butanol stripping during fermentation [7] are certainly valuable and will increase product yield, improving 1-butanol tolerance through metabolic engineering remains important to establishing a viable production strain. This research seeks to create several more 1-butanol tolerant strains and determine which phenotypic traits are closely associated with improved tolerance.

Metabolic engineering

Previously, the focus on genetic manipulations of clostridium fermentation strains has led to better understanding of the alcohol stress response, including the role of gene regulatory machinery [2]. Mutational analysis, targeted metabolic engineering, and combinatorial metabolic engineering approaches have since been used to achieve a staggering 300% increase in 1-butanol

production [8, 9] with up to 80% improved 1-butanol tolerance compared to the parent strain [10, 11]. The master regulator of solvent production and sporulation in *C. acetobutylicum*, Spo0A, was identified (along with its regulons) as a target for metabolic engineering to improve solvent production and tolerance [12]. Simultaneously, genes related to the cell stress response, including the heat shock response and chaperone proteins (i.e., *dnaK* and *groESL*) were also examined to engineer more tolerant strains of the Gram-positive *Bacillus subtilis* and *C. acetobutylicum* [10, 13-15]. Rational metabolic engineering approaches often involve targeted gene overexpression, knock-down, knockout, or mutation [4]. One approach targeted the cyclic AMP receptor protein (CRP) in a mutagenic screen and found 100% increase in butanol tolerance in *E. coil* at 1.2% butanol; affecting cell membrane adhesion, cell shape, and expression of multiple stress response genes [16]. Another approach involved expression of a reactive oxygen species scavenger metallothionein from tilapia in *E. coli* to absorb oxidative products of butanol stress and increase butanol tolerance by 50% [17].

Despite advances in tolerance through the addition and removal of individual genes and regulators, most of the cell's tolerance comes from the action of multiple genes in tandem [18, 19]. It is the expression and repression of several genes (many still unknown) which control the cellular response to alcohol exposure and determine the final cell phenotype [20]. Notably, alcohol exposure has been shown to disrupt the membrane fluidity and affect the saturated lipid content [21], leading to ATP leakage and disruption of the pH gradient [22], increased auto-lysis [23], and impaired sugar transport [24]. Reactive oxygen species have been found in *E. coli* during exposure to alcohol stress [17], and high-throughput 1-butanol and isobutanol tolerance screens have revealed the best performing microbes in soil samples often have additional or modified fatty acid synthase enzymes that modulate lipid bilayer composition [25].

Genomic libraries

While engineering individual gene targets has demonstrated substantial improvement in 1-butanol yield, a DNA library enrichment screen can be used to test millions of possible targets simultaneously. DNA library enrichment is a combinatorial approach to metabolic engineering that involves packaging DNA fragments (usually covering an entire genome) into an expression plasmid and subjecting the culture to a selective pressure. Fragments that improve microbial fitness (e.g., 1-butanol tolerance) will result in higher growth rates (or resistance to cell death) and enrichment in the culture, while cells harboring neutral or harmful fragments will remain few in the culture. This technique has been used alongside DNA hybridization microarrays to identify coding and non-coding genetic elements that aid in 1-butanol survival. In particular, a transcriptional regulator element (CAC1869) was identified that allowed *C. acetobutylicum* to survive 90% higher 1-butanol concentrations than the wild-type [26]. The multi-scale analysis of library enrichment (SCALES) approach is a platform developed to monitor the enrichment dynamics of a DNA library. A snap-shot of the plasmid-borne DNA library fragments in the cell culture were captured at various stages of enrichment under ethanol and 1-butanol stress and were hybridized to DNA microarrays with 4,200 open reading frame targets. The genes identified during the enrichment process were dedicated primarily to LPS and lipid biosynthesis, transcription and translation, fatty acid oxidation, stress response, drug resistance, serine synthesis, membrane proteins, and sugar and amino acid transport [19, 27, 28]. A similar study performed on *Saccharomyces cerevisiae* demonstrated the importance of protein degradation genes particular to the ubiquitin proteasome system for tolerating C3 and C4 alcohol stress [29]. Simultaneous identification of multiple genes imparting alcohol tolerance is important for deriving rational metabolic engineering strategies for alcohol tolerance.

While genomic library creation methods are available in the literature [30, 31], DNA library construction requires expert-level expertise in the extraction and collection of bulk genomic DNA (gDNA), processing (through physical shearing) to the desired insert size, polishing DNA ends, efficiently cloning these fragments into an expression vector, and transforming the host to produce thousands of microbial colonies per plate. We experienced low-efficiency cloning of DNA libraries produced through shearing and end-polishing gDNA, which prompted us to explore other methods of library generation. Degenerate oligonucleotide primed PCR (DOP-PCR) is another candidate technology for library generation that uses PCR primers containing random nucleotides to amplify whole genomes. In one implementation, over 80% of 144 assayed STR loci in the human genome were amplified by DOP-PCR [32]. High-throughput sequencing has since indicated shortcomings in genome coverage and amplification uniformity by DOP-PCR compared to non-PCR based approaches [33]; however, DOP-PCR provides a reliable means of generating genomic libraries from a broad host of organisms without the need for cultivation [34].

Microbial phenotyping using Raman spectroscopy

Recently, we showed that Raman spectroscopy can be used to study the changes in cell phenotypes when *E. coli* cultures are subjected to toxic levels of 1-butanol [35] and different antibiotics [36]. Raman spectroscopy is a powerful analytical technique for microbial phenotyping, with its major advantages being: (i) information is obtained in near real-time, (ii) the analysis is label-free and requires minimal or no sample preparation, and (iii) there is no spectral interference from water [37, 38]. Cells having unique Raman signals from macromolecules such as nucleic acids, proteins, lipids, and carbohydrate polymers can be used to

further generate "chemometric fingerprints" of whole cells, which can be compared to determine similarity of expressed phenotypes [39, 40]. Different phenotypes have been investigated (employing Raman spectroscopy) in the literature in relation to: (i) solvent tolerance [41-43], (ii) antibiotic mechanism of action [36], bacterial ecology, and evolution [44]. In our recent work, we identified individual Raman bands that described the phenotypic changes of *E. coli* cultures in response to 1-butanol toxicity [35]. In particular, Raman band intensities were correlated with traditional analytical methods (e.g., GC-MS/FID) for (i) saturated (Raman intensity (I) at wavenumber 2870 cm^{-1}), (ii) unsaturated (I1263 cm^{-1}), and (iii) cyclopropane fatty acids (I1554 cm^{-1}). Membrane fluidity was found as a ratio of Raman band intensities (I2852/I2924), and Raman bands were identified to reveal total protein content as well as amino acids composition of *E. coli* cells. These individual Raman bands and the technique of chemometric fingerprinting using entire Raman spectra were used in this research for phenotyping 1-butanol tolerant *E. coli* strains found through genomic library enrichment.

Phenotyping 1-butanol tolerant strains

The DOP-PCR approach was used in this research to generate genomic DNA libraries from *E. coli*, *Clostridium acetobutylicum*, and environmental metagenome DNA sources. These libraries were cloned and introduced into *E. coli* and enriched under selective pressure from toxic levels of 1-butanol. Strains more tolerant to 1-butanol were obtained, and Raman spectroscopy was used to characterize the physiological differences between the tolerant strains with and without 1-butanol exposure. Here, we report for the first time that several different *E. coli* phenotypes are capable of conferring tolerance to toxic levels of 1-butanol. This suggests multiple routes to 1-butanol tolerance are possible.

MATERIALS AND METHODS

Strains and cultivation

E. coli NEB 10-beta (New England Biolabs; Ipswich, MA) and *C. acetobutylicum* ATCC 824 were cultivated and maintained on Luria-Burtani (LB) media and 2xYTG media (pH 5.8; 16 g Bacto tryptone, 10 g yeast extract, 4 g NaCl, and 5 g glucose per L) respectively. Both were grown in liquid media and on solid agar plates. *E. coli* were grown aerobically at 37°C. *C. acetobutylicum* was grown anaerobically at 37°C in an $N_2/CO_2/H_2$ (90/5/5%) environment in a Bactron anaerobic chamber (Shel Lab; Cornelius, OR). Culture growth in liquid media was monitored by optical density at 600 nm (OD_{600}).

Isolation of template genomic DNA

E. coli gDNA was extracted using a Generation Capture column (Qiagen; Valencia, CA), and plasmid DNA was extracted using a GeneJET miniprep kit (ThermoFisher Scientific; Waltham, MA) following the manufacturer's protocol. *C. acetobutylicum* gDNA was extracted as described previously [45]. Metagenomic DNA was extracted from a 5 g soil sample isolated from the Stroubles steam bank on the Virginia Tech campus in Blacksburg, Virginia (USA). For this sample, gDNA was extracted using the QIAmp® DNA Stool Mini Kit (Qiagen) with the following modifications. Briefly, 5 g of unprocessed soil was mixed with 5 mL of deionized water for 5 min and incubated at 95°C for 10 min. The sample was then centrifuged at 3000 rpm, and the supernatant was added to a second tube and mixed with the kit-supplied ASL Buffer and 1 InhibitEX™ tablet for 1 min at room temperature. This sample was centrifuged at 5000 rpm, combined with 1 volume of kit-supplied AL Buffer and 1 volume of absolute (100%) ethanol.

The sample was then applied to the QIAmp spin column, washed, and eluted following manufacturer's protocol.

DNA library generation by DOP-PCR

DNA libraries were generated by DOP-PCR [32] from gDNA templates from (i) *E. coli* NEB 10-beta (a DH10B derivative), (ii) *C. acetobutylicum* ATCC 824, and (iii) an environmental soil metagenomic sample. DOP-PCR DNA library construction used the PCR primer 9NATG (5'-TAG ACA ATG GGG CAT NNN NNN NNN ATG-3') (Integrated DNA Technologies; Coralville, IA). The 9NATG primer contains degenerate nucleotides, specified by "N," that have equal probability of being A/T/C/G and was synthesized as a random mixture. This approach allowed for partial and mismatched annealing with random segments of the template gDNA. The DOP-PCR reaction was run with the 9NATG primer and *Taq* polymerase. gDNA template was diluted to 10 ng/µL, and 1 µL of this mixture was combined with 5 µL of 10x ThermoPol® Reaction Buffer (20mM Tris-HCl, 10mM $(NH_4)_2SO_4$, 10mM KCl, 2mM $MgSO_4$, 0.1% Triton®X-100), 1µL 9NATG primer (100uM), 1.25 µL dNTP (10mM each), 1 µL 5U/µL NEB *Taq* Polymerase, and 42 µL of chilled Type I ultra-pure water. The *Taq* polymerase was used for DNA library generation since it leaves an adenine (A) overhang on PCR products. This was found to facilitate cloning into a linearized T-tailed vector, as discussed below. The DOP-PCR reaction was carried out on an Eppendorf gradient MasterCycler PCR with the following program: (i) 6 min initial denaturation at 95°C, (ii) 1 min denaturation at 95°C, (iii) 3 min anneal at 30°C, (iv) 3 min ramp to 72°C, and a (v) 3.5 min elongation at 72°C. Steps (ii) through (v) were repeated 10 times. The low stringency cycle was followed immediately by a second high stringency cycle: (vi) 1 min denaturation at 95°C, (vii) 1 min annealing at 55°C,

(viii) 3 min elongation at 72°C, and (ix) a final elongation for 10 min at 72°C. In this program, steps (vi) through (viii) were repeated for 25 cycles. The resulting DOP-PCR generated DNA library was visualized on an ethidium bromide Tris-Acetate agarose gel. DNA above 1000 bp was excised and purified in a GeneJet spin column (ThermoFisher Scientific) for cloning.

DNA library cloning

A total of 50 ng of the linearized T-tailed vector pCR8/GW/TOPO TA (Life Technologies; Grand Island, NY) was combined with 0.5 µL of the purified DOP-PCR amlpified DNA mixture (50 ng/µL) according to manufacturer's directions and incubated at room temperature for 1 hour. Transformation of 2 µL of the pCR8 ligation into 20 vials of 50 µL of Stellar® chemically competent *E. coli* (Clontech; Mountain View, CA) generated 40,000-60,000 colonies as verified by LB agar plates for each DNA library transformed. The remaining transformed cultures were pooled and grown in 200 mL of liquid LB with 50 mg/L spectinomycin to an OD_{600} of 0.6. Cultures were then centrifuged in 50 mL aliquots for storage at -80°C and for plasmid extraction using the GeneJet Midiprep kit (ThermoFisher Scientific). Following this amplification step, the DNA library was transferred from the entry vector to a suitable destination vector for expression.

Transfer of the DNA library between the pCR8/GW/TOPO TA entry vector and the pDEST^{TM}14 (Invitrogen) destination vector was done by a recombination reaction using LR Clonase II (Invitrogen) according to the manufacturer's protocol. Aliquots of 1 µL transformation mix were used to transform ten 50 µL vials of NEB5α Competent *E. coli* (New England Biolabs) to maximize recombinant transformants from the costly LR ClonaseII reaction. The resulting transformants were pooled once again, generating in excess of 100,000 colonies on

LB agar plates with ampicillin (100 mg/L). Cells were grown to OD600 1.5 in 200 mL of liquid LB and the pDEST library plasmids were harvested using the alkaline lysis procedure[46]. Bulk plasmid was introduced into NEB *E. coli* ER2556 made chemically competent by the $CaCl_2$ method [47] to generate $>1x10^6$ colonies.

The plasmid control (pCTL) was generated by digesting plasmid pDEST^{TM}14 with the enzyme EcoRI (New England Biolabs) and re-ligating, forming a plasmid with no Gateway cloning cassette. This strategy disrupted the chloramphenicol resistance gene and removed the *ccdB* cell death selection gene as well as the T7 terminator. The T7 promoter remained un-induced in all trials.

Genomic coverage analysis

Size-selected DOP-PCR generated DNA fragments were submitted for Mi-Seq 150 bp paired end sequencing at the Virginia Bioinformatics Institute on the Virginia Tech campus. Data were analyzed using freely available software as described below. Sequencing reads were filtered to remove the 5' 9NATG primer ends and adapters used in sequencing using the Fastx-Toolkit [48] and Trimmomatic [49]. For *E. coli* and *C. acetobutylicum* DNA libraries, cleaned fragments were aligned to the respective genomes using Bowtie2 v 2.1.0 [50] and Samtools v 1.0 [51], and coverage analysis was performed using Qualimap v2.0 [52]. The environmental metagenomic DNA library was analyzed through the MG-RAST data server [53] under accession number 4583655.3. The 5' ends of paired end reads were trimmed of residual 9NATG non-template sequences and paired ends were combined into single data set after upload. Quality and duplicate trimming were disabled to analyze the effect of the DOP method on read bias.

1-Butanol challenge

Approximately 40-60 x10^3 mutants for each library (i.e., *E. coli*, *C. acetobutylicum*, and environmental metagenome) were pooled in separate cultures and used to inoculate 100 mL of liquid LB media (containing 100mg/L ampicillin) and incubated at 37°C and 250 rpm. Upon reaching an OD_{600} of 0.8, cultures were used as inoculum for liquid LB (100mg/L ampicillin) with 0.5% (v/v) 1-butanol (Sigma-Aldrich; St. Louis, MO) in a 500 mL sealed bottle to prevent butanol evaporation. Upon reaching an OD_{600} of 0.8, samples were used to inoculate media at 0.9% (v/v) 1-butanol, followed by 1.1%, and 1.3% 1-butanol at which no change in culture turbidity was noticed after 24 hours. Surviving samples at 1.3% butanol were plated on LB agar with 100 mg/L ampicillin and mutants from each plate were isolated and sequenced using the universal T7 promoter and terminator primers 5'-TAATACGACTCACTATAGGG-3' and 5'-GCTAGTTATTGCTCAGCGG-3'. Plasmid DNA was extracted from each isolate and introduced into chemically competent *E. coli* NEB ER2566 expression cells. After an overnight outgrowth, each sample was diluted to an OD_{600} of 0.2 and equal volumes were used to inoculate fresh liquid LB media (containing 100 mg/L ampicillin) with 1-butanol concentrations of 0, 0.9, and 1.1% (v/v) (3-4 replicates per sample per level) in a covered Costar® 96-well plate, sealed to prevent evaporation. Samples were grown at 35°C and medium shaking, and OD_{600} was monitored every 20 minutes in a Synergy H4 plate reader (Biotek; Winooski, VT). Sample growth rates and maximum OD_{600} readings were calculated using GroFit [54]. Percent survivability was measured as the ratio of the average maximum growth rate at 0.9% or 1.1% (v/v) 1-butanol to the average maximum growth rate in unchallenged culture for each mutant.

Raman spectroscopy

In order to prepare cells for Raman scanning, an overnight culture of each mutant was used to inoculate 10 mL of liquid LB (containing 100 mg/L ampicillin). After reaching an OD_{600} of 0.5-0.7, cells were divided into 6 aliquots. Three were spiked with 1.1% (v/v) 1-butanol and incubated with shaking for 3 hours, and the other three were incubated as controls (i.e., no 1-butanol). OD_{600} measurements were taken for all samples and cultures were centrifuged to pellet cells. Supernatant was discarded, and the recovered cells were washed three times in phosphate-buffered saline. Then, 20 µL of washed cells from each sample were air dried on an aluminum surface at room temperature. The dried cells were analyzed by Raman spectroscopy. Raman scans were acquired using a Bruker Senterra dispersive Raman spectrometer equipped with a confocal microscope and objective lens of 100x magnification (Bruker Optics, Billerica, MA). The following instrument settings used to acquire scans: 532 nm laser excitation, 20 mW laser power, 10 s exposure time, and a spectral resolution of 9-15 cm^{-1}. Each sample was scanned a minimum of 6 times prior to data pre-processing (baseline correction) using the OPUS software (Bruker Optics) and further analysis using custom scripts written in MATLAB (R2012A) (MathWorks, Natick, MA). Two types of analysis were performed: (i) Raman band intensity analysis following vector normalization of spectra [35] and (ii) chemometric fingerprinting of whole Raman spectra using principle component analysis and linear discriminant analysis[36]. Both of these analysis techniques have been described in the literature.

RESULTS

DNA library production and enrichment

DNA libraries were produced using the DOP-PCR method and cloning techniques described above. Separate libraries were produced from different gDNA templates: (i) *E. coli* NEB 10-beta, (ii) *C. acetobutylicum* ATCC 824, and (iii) the environmental soil metagenome. Results of DNA library generation are shown in Figure 3-1. Details of the *E. coli* and *C. acetobutylicum* DNA libraries are shown in Figure 3-1a and provide the percent coverage and times coverage parameters for each library. The percent coverage describes what percentage of the native genome is present in the DNA library. The times coverage describes the abundance of DNA library fragments. For example, the *E. coli* DNA library covers 92% of the *E. coli* genome at least one time and it covers 56% of the genome at least 10 times. In addition, 24% of the genome is covered at least 50 times in the library. Furthermore, the *C. acetobutylicum* DNA library yielded 28% genome coverage at least 1 time, with 10% of the genome being covered at least 9 times. This points to very different amplification properties of the 9NATG primer in the DOP-PCR protocol and suggests future optimization is warranted when dealing with genomes with GC content (i) greater than 50%, (ii) nearly equal to 50% (i.e., *E. coli*, 50.8%), and (iii) less than 50% (i.e., *C. acetobutylicum*, 30.9%). The previous chapter, focusing on optimization of the DOP-PCR reaction, demonstrates substantial improvement in the genomic coverage of the low GC species. In particular, thermodynamically constrained primer designs (Mset and L1 design) enabled amplification of 88% of the *C. acetobutylicum* genome at least once. However, only the 9NATG library was assembled at the time of this study.

The 9NATG primer was used with the environmental metagenome sample in a DOP-PCR procedure to produce the environmental DNA library. The metagenomic DNA fragments

that matched to targets on the MG-RAST server with a maximum e-value cutoff of 10^{-5} and minimum identity of 60% resulted in a metagenome GC content of 53±10% and represented 4773 organisms and 569220 identifiable protein coding regions as shown in Figure 3-1b. The DNA library includes DNA fragments from 72 phyla (listed in Figure 3-1). The Proteobacteria made up the largest fraction of the library (67%) with the Bacteriodetes making up the next largest fraction (9.3%). The Cyanobacteria accounted for 1.4% of the DNA library, and 65 of the 72 phyla contributed less than 1%. The "number of library fragments" in Figure 3-1b refers to the number of DNA fragments identified through sequencing, not the number existing in the entire library.

The *E. coli* (ECO), *C. acetobutylicum* (CAC), and environmental soil metagenome (ENV) DNA libraries were effectively cloned and were enriched in *E. coli* under 1-butanol stress to identify fragments that confer 1-butanol tolerance. The T7 promoter of the pDEST^{TM}14 plasmid was induced and un-induced in preliminary studies, and superior results were found from the un-induced configuration (data not shown), meaning that DNA fragment expression was dependent on native promoter activity. This approach has been implemented elsewhere as well [55]. A total of 9 DNA fragment bearing plasmids (three from each DNA library) were enriched in separate experiments. The plasmids were isolated and enriched in the majority of colonies after plating. The enriched DNA fragments were sequenced and are summarized in Table 3-1. Fragments containing the upstream region of a gene are expected to enable DNA transcription from the native promoter. Fragments containing no upstream leading DNA may contain internal promoters possibly generating regulatory RNA (or sRNA) that impacts native gene translation. Sample ENVG (Table 3-1) could not be aligned to known sequences and was characterized using BLASTx [56].

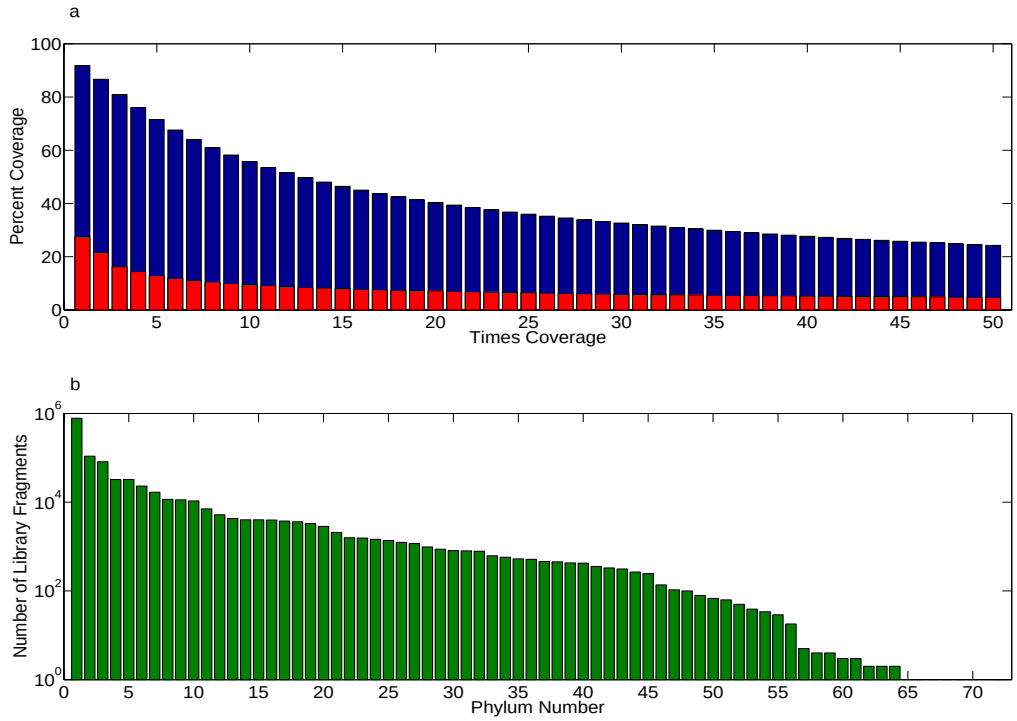


Figure 3-1. (a) The percent coverage and the number of times of genome coverage for the *E. coli* library (blue) and the *C. acetobutylicum* library (red). (b) Composition of the environmental metagenome library. The following phyla are included (as numbered): (1) Proteobacteria, (2) Bacteroidetes, (3) Verrucomicrobia, (4) Firmicutes, (5) Planctomycetes, (6) Actinobacteria, (7) Cyanobacteria, (8) Acidobacteria, (9) unclassified (derived from Viruses), (10) Chloroflexi, (11) Chlorobi, (12) unclassified (derived from Eukaryota), (13) Nitrospirae, (14) unclassified (derived from Bacteria), (15) Euryarchaeota, (16) Streptophyta, (17) Deinococcus-Thermus, (18) Chordata, (19) Spirochaetes, (20) Ascomycota, (21) Gemmatimonadetes, (22) Lentisphaerae, (23) Chlamydiae, (24) Arthropoda, (25) Aquificae, (26) unclassified (derived from unclassified sequences), (27) Thermotogae, (28) Chlorophyta, (29) Fusobacteria, (30) Cnidaria, (31) Synergistetes, (32) Bacillariophyta, (33) Deferribacteres, (34) Nematoda, (35) Basidiomycota, (36) Phaeophyceae, (37) Apicomplexa, (38) Crenarchaeota, (39) Chrysiogenetes, (40) Poribacteria, (41) Fibrobacteres, (42) Tenericutes, (43) Dictyoglomi, (44) unclassified (derived from other sequences), (45) Elusimicrobia, (46) Echinodermata, (47) Placozoa, (48) Thaumarchaeota, (49) unclassified (derived from Archaea), (50) Hemichordata, (51) Platyhelminthes, (52) Microsporidia, (53) Korarchaeota, (54) unclassified (derived from Fungi), (55) Mollusca, (56) Euglenida, (57) Porifera, (58) Annelida, (59) Rotifera, (60) Blastocladiomycota, (61) Neocallimastigomycota, (62) Brachiopoda, (63) Nanoarchaeota, (64) Xanthophyceae, (65) Chytridiomycota, (66) Entoprocta, (67) Eustigmatophyceae, (68) Glomeromycota, (69) Haplosporidia, (70) Rhombozoa, (71) Tardigrada, and (72) Thermodesulfobacteria.

Table 3-1. Description of nine DNA fragments enriched from *C. acetobutylicum* (824), *E. coli* (ECO) and environmental soil sample (ENV) DNA libraries.

Mutant Name	**Gene Match**	**Gene Annotation**	**Genomic Context of DNA Library Fragment** ***Green arrows: Library insert*** ***Red arrows: Reference Genes (5kb window)***
824A	CAC2147 (*flhA*)	Flagellar biosynthesis protein	
824C	CAC1044	NDH:Flavin oxidoreductase, possible NADH oxidase	
824D	CAC1866 CAC1867 CAC1868	Putative 4-Cys Ferrodoxen; XRE family transcriptional regulator (phage-related); Uncharacterized secreted protein	
ENVE	*Methanosalsum zhilinae*[1] Mzhil_1433	Restriction modification	N/A
ENVF	*Desulfobulbus propionicus*[1] Despr_1510	Sulfatase	N/A
ENVG	*Candidatus entotheonella*[1] sp. TSY2	Amidohydrolase protein domain	N/A
ECOH	ECDH10B1295 (*rssB*) ECDH10B1296 (*galU)*	Response regulator of RpoS; UTP glucose 1-phosphate uridylyltransferase	
ECOI	ECDH10B0351 (*maK*) ECDH10B0352 (*araJ*)	Manno(fructo)kinase; Putative transport protein	
ECOJ	ECDH10B3103 (*yggC*) ECDH10B3105 (*yggD*) ECDH10B3106 (*yggF*)	Predicted nucleoside triphosphate hydrolase; Predicted transcriptional regulator (fructose transport); Fructose 1,6-biophosphase	

[1]DNA sequences matches were not available. Samples were matched using BLASTx to find protein domain similarities and the most probable host is listed.

The DNA fragment containing plasmids were re-transformed into the *E. coli* ER2556 cell line, which was grown in the presence of 0%, 0.9%, and 1.1% (v/v) 1-butanol for further analysis. Results of maximum culture growth rate and maximum OD_{600} are shown in Figure 3-2 for cultures containing each of the 9 enriched DNA fragments, the pCTL plasmid control, and the *E. coli* ER2556 grown without a recombinant plasmid and antibiotics in the medium. Merely the presence of different DNA fragments altered the maximum observable growth rate and OD_{600} in cultures void of 1-butanol (Figure 3-1). To determine the ability of each DNA fragment to confer 1-butanol tolerance, the maximum growth rate of each strain at 0.9% and 1.1% (v/v) 1-butanol was divided by the value obtained in absence of 1-butanol. This was referred to as the "cell survivability ratio" with respect to the maximum growth rate. These values were found to be statistically different compared to the plasmid control (pCTL) in 5 out of the 9 strains tested. Calculations were performed using the Multiple Comparison Test in MATLAB with a 95% confidence level.

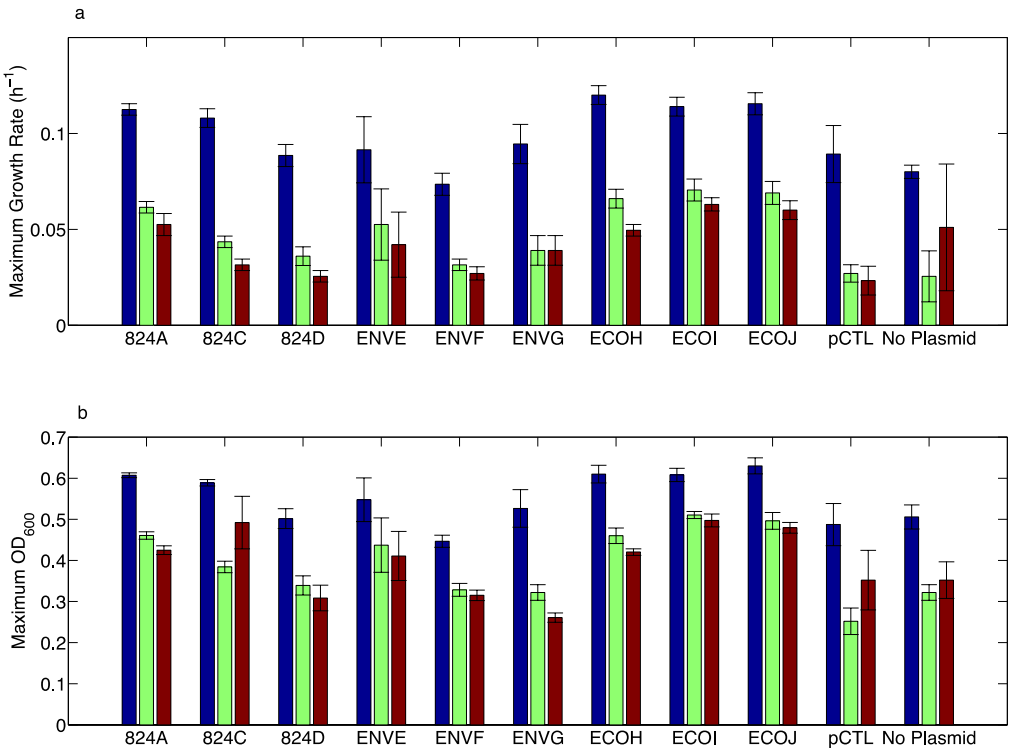


Figure 3-2. (a) Maximum growth rate (h-1) and (b) maximum OD600 for strains harboring enriched library inserts, the control plasmid (pCTL), and cells harboring no plasmid ("No Plasmid"). Results are shown for growth in absence of butanol (blue), 1.1% (v/v) 1-butanol (green), and 0.9% (v/v) 1-butanol (red).

Furthermore, the percent difference of these ratios relative to the plasmid control (pCTL) strain was calculated. The cell survivability ratio with respect to the maximum OD_{600} was also calculated using the same method. These results are given in Table 3-2 and describe 1-butanol tolerance (in maximum growth rate and OD_{600}) of the engineered strains relative to the pCTL control. From these calculations, the strain containing the ECOI DNA fragment showed significantly less growth reduction with 1-butanol exposure. The maximum growth rate reduction was improved by over 100% compared to the pCTL control strain, and the reduction in

maximum OD_{600} was improved by 62% in 0.9% (v/v) 1-butanol and 13% in 1.1.% (v/v) 1-butanol. All strains showed a higher cell survivability ratio with respect to maximum growth rate in both 0.9% and 1.1% (v/v) 1-butanol, which is likely why these DNA fragments were selected during library enrichments. In addition, the cell survivability ratio with respect to maximum OD_{600} was significantly higher at 0.9% (v/v) 1-butanol, but these values reduced significantly for the 1.1% (v/v) 1-butanol treatment. This is likely because of the relatively high reading obtained for the pCTL plasmid control (Figure 3-2b).

Table 3-2. Increases in cell survivability ratios (with respect to maximum growth rate and OD600) relative to the plasmid control (pCTL) for growth in 0.9% and 1.1% (v/v) 1-butanol.

Cell Survivability Ratio With Respect To:	824A	824C	824D	ENVE	ENVF	ENVG	ECOH	ECOI	ECOJ
Max Growth Rate 0.9% (v/v) 1-butanol	81%	33%	34%	90%	42%	36%	82%	104%	97%
Max Growth Rate 1.1% (v/v) 1-butanol	79%	12%	11%	76%	41%	58%	58%	112%	99%
Max OD_{600} 0.9% (v/v) 1-butanol	47%	26%	31%	54%	42%	18%	46%	62%	52%
Max OD_{600} 1.1% (v/v) 1-butanol	-3%	16%	-15%	4%	-2%	-31%	-5%	13%	5%

Raman spectroscopy

Cultures harboring plasmids containing the DNA fragments, and controls, were analyzed by Raman spectroscopy. The spectra were baseline corrected, vector normalized, and plotted in Figure 3-3. Spectra of all plasmid-containing cultures are shown in Figure 3-3a (with 1-butanol treatment) and Figure 3-3b (no 1-butanol treatment). The 1-butanol treated and un-treated spectra

for the control strain with no plasmid (or antibiotic) are shown in Figure 3-3c and clearly show differences from the spectra of plasmid-containing cultures. Thus, all comparisons for strains containing enriched DNA fragments were made relative to the pCTL plasmid control, which had a more similar phenotype due to the presence of ampicillin in the culture media. With 1-butanol treatment (Figure 3-3a), the Raman spectra of the DNA fragment containing strains departed significantly from the pCTL plasmid control strain (thick blue line in Figure 3-3a). In particular, the most significant deviations occurred between 800-1000 cm^{-1}. Several functional groups of macromolecules have been assigned to individual Raman bands in this region [57]. The following are a subset with potential relevance to 1-butanol induced phenotypes: phosphodiesters (825 cm^{-1}), glycogen (853 cm^{-1}), proline and tyrosine side chains (855 cm^{-1}), C-C skeletal backbone (893 cm^{-1}), carbohydrates (931 cm^{-1}), α-helix structures (951 cm^{-1}), lipids (968 cm^{-1}), =CH bending (980 cm^{-1}), and phenylalanine (indicative of protein) (1002 cm^{-1}). However, it is noted that Raman band assignments are often system-dependent. The comparison of Raman spectra in cultures un-treated with 1-butanol do not show the strong deviation from 800-1000 cm^{-1}. Instead, variation is seen in other regions such as 1400-1500 cm^{-1} and 1600-1700 cm^{-1}, indicating that the mere presence of the DNA fragments (without 1-butanol) is enough to induce phenotypic changes in the cultures that are detectable by Raman spectroscopy.

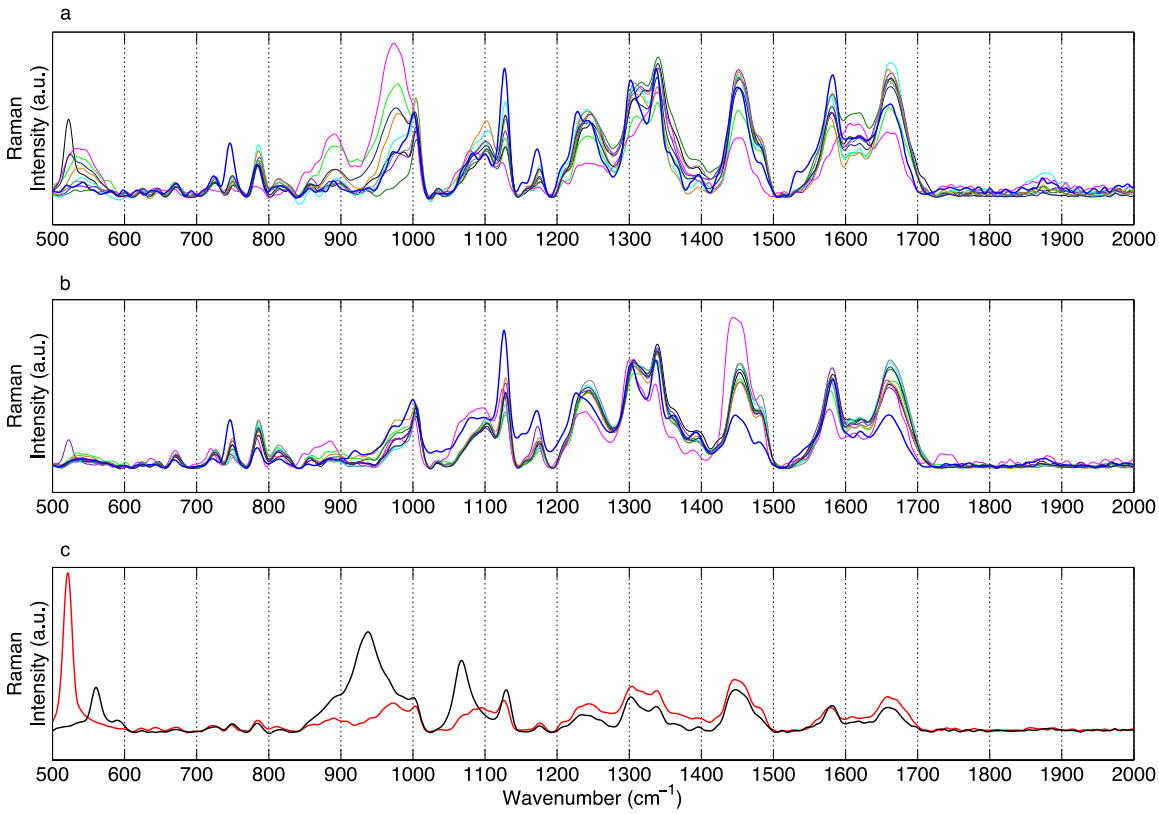


Figure 3-3. Raman spectra acquired of the following: (a) cultures containing library inserts with 1-butanol exposure, (b) cultures containing library inserts with no exposure, and (c) wild-type cells with 1-butanol exposure (red) and without exposure (black). The following are shown in (a) and (b): 824A (black), 824C (bright green), 824D (dark green), ENVE (cyan), ENVF (magenta), ENVG (violet), ECOH (orange), ECOI (grey), ECOJ (dark blue), and the pCTL control (blue – thick line)

The analysis of individual Raman band intensities was conducted based on the sets of Raman bands identified in our previous research [35]: saturated (I2870 cm^{-1}), unsaturated (I1263 cm^{-1}), cyclopropane fatty acids (I1554 cm^{-1}), and membrane fluidity (I2852/I2924). In addition, nucleic acids were measured at I1576 cm^{-1} and phenylalanine (indicative of total protein) at 1002 cm^{-1} [57]. This enabled calculation of the total fatty acids to nucleic acids ratio and total fatty

acids to phenylalanine (protein) ratio. These results are shown in Figure 3-4 for the 1.1% (v/v) 1-butanol treated cultures harboring plasmids with the enriched DNA fragments. Results are presented as percent changes relative to the pCTL control plasmid strain. All strains with enriched DNA fragments showed increases in unsaturated fatty acids, membrane rigidity, and the ratio between total fatty acids to nucleic acids. Likewise, all cultures showed a decrease in cyclopropane fatty acids. However, only 4 of 9 cultures showed increases in saturated fatty acid content and the ratio of total fatty acids to phenylalanine (protein) showed mixed results as well. This indicates that the enriched DNA fragments induced different 1-butanol tolerance mechanisms in response to stress.

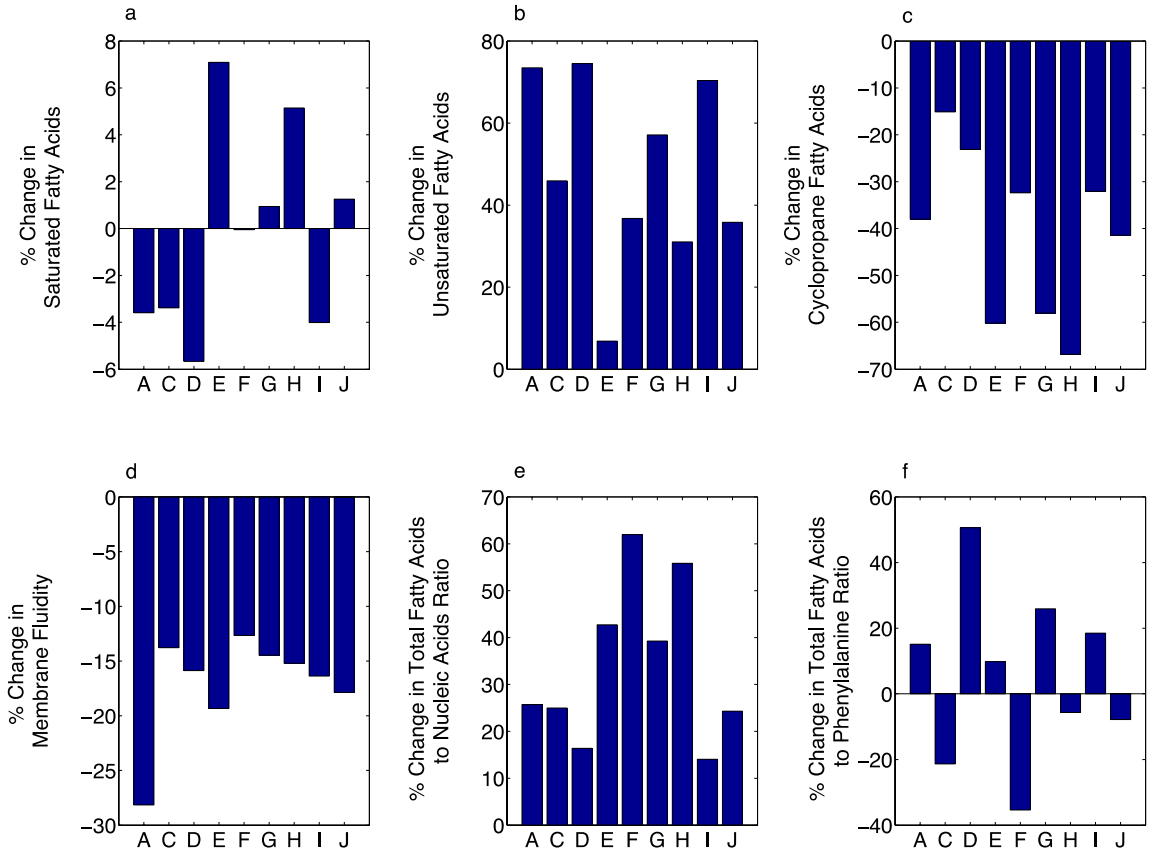


Figure 3-4. Raman spectroscopy phenotyping results for cultures exposed to 1.1% (v/v) 1-butanol. The percent composition change for (a) saturated, (b) unsaturated, and (c) cyclopropane fatty acids for strains harboring enriched library fragments. (d) The percent change in membrane fluidity, (e) thepercent change in ratio of total fatty acids to nucleic acids, and (f) the percent change in the ratio of total fatty acids to phenylalanine (i.e., protein). All percent change values are relative to the plasmid control strain (pCTL). The following library inserts are shown: (A) 824A, (C) 824C, (D) 824D, (E) ENVE, (F) ENVF, (G) ENVG, (H) ECOH, (I) ECOI, and (J) ECOJ.

Chemometric fingerprinting

Another method of comparing phenotypes is chemometric fingerprinting of whole Raman spectra. In this analysis, the band intensities of entire Raman spectra (as opposed to a few selected bands) are used for comparisons. The technique uses linear discriminant analysis and has been described elsewhere [36]. Chemometric finger printing was used in this research for two purposes: (i) determine the similarity between strains expressing enriched DNA fragments and (ii) probe for differences between 1-butanol treated and untreated cultures. Results are shown in Figure 3-5. In the first application (Figure 3-5a), Raman spectra of 1-butanol treated cultures were analyzed, and several unique phenotypes were realized as all samples separated from one another on a canonical plot, where clustering indicates similarity of phenotypes. Importantly, the pCTL phenotype was separated from all other cultures, meaning all DNA fragments led to a significant change in phenotype. Phenotypes of 824C and ENVF separated significantly from the others. Both of these phenotypes showed decreased total fatty acids to phenylalanine (total protein) ratio in Figure 3-4f. In addition, the phenotype of 824A also showed significant separation from the others, and it had the greatest increase in membrane rigidity (Figure 3-4d). Furthermore, strains originating from the *E. coli* genomic library (ECOH, ECOI, and ECOJ) clustered somewhat closely together. Only with DNA from other genomes were substantially different phenotypes observed (824A, 824C, ENVE, ENVF). However, ECOI was identified in Table 3-2 as having the greatest increase in 1-butanol tolerance. Its closest phenotypes (824D and ECOH) did not show the same increase in tolerance. When all Raman spectra (from all strains) were considered simultaneously and discriminated according to 1-butanol treatment, clear separation resulted (Figure 3-5b). This demonstrates that regardless of DNA fragment, phenotypes were largely induced by 1-butanol treatment.

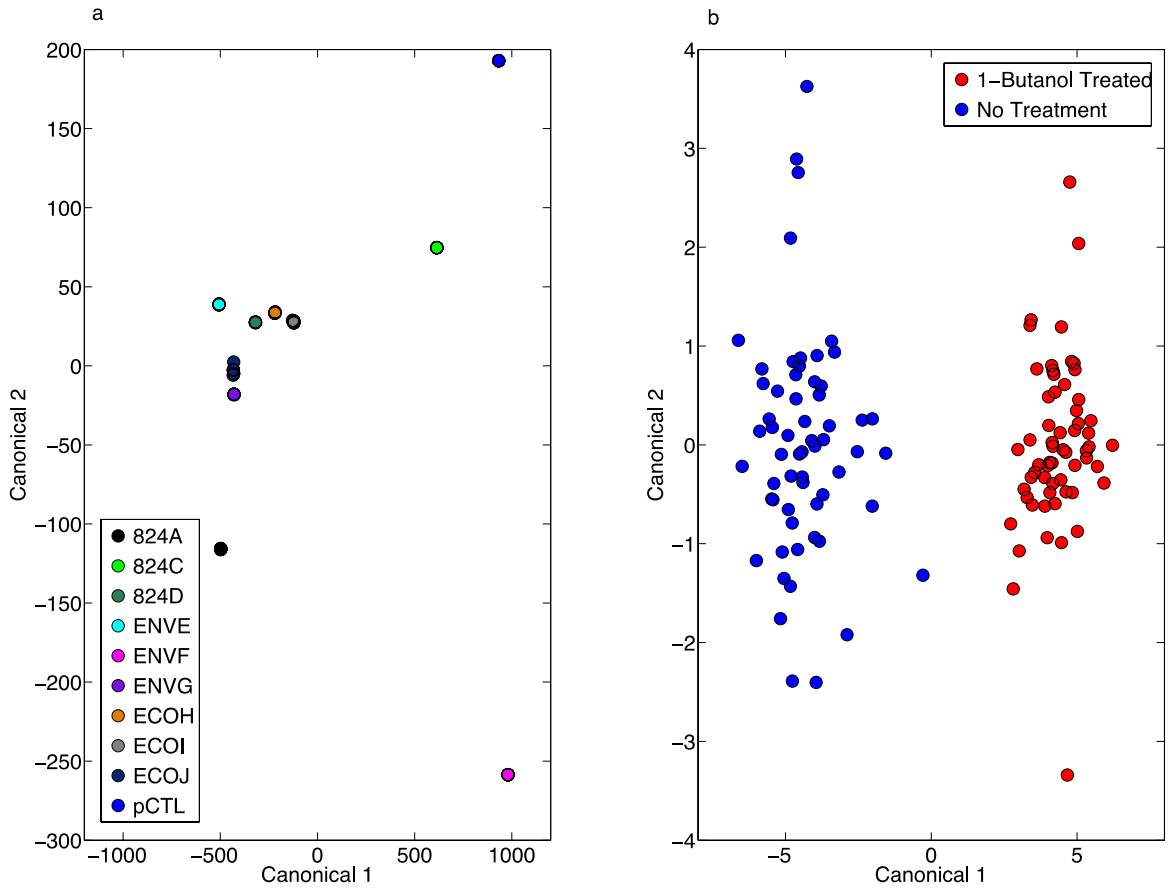


Figure 3-5. Results of Raman spectroscopy chemometric fingerprinting. (a) Classification of strains exposed to 1.1% (v/v) 1-butanol according to the library insert contained on a plasmid and (b) classification of all samples according to 1-butanol exposure (treated or not treated).

DISCUSSION

Advantages of the DOP-PCR approach

A new method of DNA library generation through DOP-PCR was examined in this research along with the phenotypic characterization of 1-butanol tolerant strains resulting from library enrichment. The effectiveness of generating DNA libraries using DOP-PCR was investigated using next-generation sequencing. The libraries generated using *E. coli* NEB 10-β

gDNA showed greater than 90% coverage at the 1 time coverage level. This number is likely an underestimate due to the 113-kb region of the genome that exists as a tandem repetitive genetic element. As the Bowtie2 aligner cannot distinguish between the two duplicate regions, reads aligning to either region are assigned only to the first, explaining up to 2.5% of the total lack in genomic coverage [58]. The amplification technique was also applied to the *C. acetobutylicum* ATCC 824 genome; however, much lower coverage of the reference genome was achieved (Figure 3-1a). These results may be indicative of the much lower GC content of the genome and the presence of the native pSOL1 megaplasmid which has been shown to bias whole genome amplification [33, 59]. However, even with an admittedly limited sample, the enriched DNA library contained fragments adjacent to identified XRE family regulators in earlier genetic screens [60], the enriched sugar transporter (*rrsb*), and flagellar proteins. These were identified in 4 of the 9 enriched DNA fragments, and they have also been identified as important contributors in DNA microarray and targeted gene knockout studies [28, 61]. Clearly, there is room for DOP-PCR library generation improvement through better primer design, but this research serves as the first demonstration of DOP-PCR amplified gDNA as a source for the creation of a DNA library. The benefits of this method enable DNA library creation from trace amounts of genetic material, potentially down to the level of single cells without the need to propagate cultures in the laboratory [62]. Further, the DOP-PCR method requires no digestion, end-polishing, or ligation steps when used with compatible linearized T-tailed vectors, allowing single-step incorporation of DNA into a plasmid library. The DOP-PCR DNA library generated from gDNA extracted from a stream bank metagenome generated a majority of fragments from the Proteobacteria and Bacteroidetes, with traces of Archaea and eukaryotic origin. Though the use of genetic amplification techniques on environmental samples is not new [63], this research

is the first to our knowledge to apply these techniques to generate DNA libraries for strain enrichment experiments.

More about enriched DNA fragments

The enriched DNA fragments (Table 3-1) align with previously published transcriptomics data describing the responses of both *E. coli* and *C. acetobutylicum* to 1-butanol stress. However, many of the enriched DNA fragments encode for partial genes, intergenic regions, and/or antisense gene fragments, so the exact mechanism(s) of how the enriched DNA fragments are conferring 1-butanol tolerance is not known at this time. Previous research found sugar transport genes to be enriched under 1-butanol stress [28], and the *araJ,* and *yggC* (also known as *frcK*) genes for *E. coli* arabinose transport and fructose transport kinase were enriched in the DNA library in this research. The *araJ* gene resides in the arabinose operon and encodes for a membrane protein suspected of being involved in the transport or processing of arabinose-containing polymers [64, 65]. The *yggD* gene is uncharacterized in *E. coli* DH10β, but it encodes for a transcriptional regulator in *E. coli* B strains [66] that represses genes related to mannitol transport (*mt1A* and *mt1D*) [67]. Drug resistance genes were also enriched in the DNA library exposed to 1-butanol stress, which is also in agreement with previous studies [28]. The enriched strains 824D and ENVG of this research contain drug resistance genes. The 824D strain contains CAC1867, a XRE family regulator known to activate in the presence of xenobiotics [68]. Interestingly, an XRE response element, CAC1869 (in the same operon as CAC1867) was enriched in a similar 1-butanol challenge of *C. acetobutylicum* [60]. Members of the amidohydrolase family of proteins like that of ENVG typically break nitrogen carbon bonds and can be found in proteins synthesizing glutamate, aspartic acid, and biotin. The hydrolase action

could also be selected for if it were able to degrade β-lactams in the media as it could increase fitness of the mutant in ampicillin-containing growth media. A global stress response gene, *rssB*, was also enriched in the ECOH strain. This gene is a regulator of RpoS, which has been found to increase culture viability at the expense of growth rate when silenced [69]. The same previous study demonstrated that the loss of the flagellar synthesis regulator gene, *fliA*, increased stress survivability. In this research, an antisense fragment of the flagellar biosynthesis gene, *flhA*, in *C. acetobutylicum*; was enriched. A promoter was not associated with this enriched fragment, suggesting a portion of it could have been transcribed as antisense RNA or sRNA. Strain ENVE contained an insert enriched from the environmental metagenome sample contained a DNA sequence similar to that of a restriction modification system from *Methanosalsum zhilinae* (a methanogen of Archaea). Unregulated restriction modification enzymes have been linked to programed cell death and may be responsible for premature activation of the cellular SOS response, similar to 1-butanol stress. The process terminates with cell membrane disruption and protein leakage [70]. Enriched fragments may silence biochemical salvage functions, either through sulfatase activity (present in strain ENVF) or by NADH:Flavin oxioreductase (CAC1044 in strain 824C) activity. While the enriched DNA fragments showed consistency with previous findings, the goal of this research was to examine the resulting phenotypes (i.e., biochemical composition of cultures). High-throughput methods of discerning the chemical mechanisms of enriched DNA fragments are certainly needed, but this technology is likely years away.

Incorporating Raman spectroscopy with DNA library enrichment

Raman spectroscopy has many advantages for culture phenotyping that lend well to high-throughput experiments. In particular, analysis can now be automated in 96-well plates in

modern instruments, and scanning takes on the order of minutes per microbial colony. While analysis times on the order of seconds are preferred, this is a significant advantage over traditional methods of analysis (e.g., GC-MS/FID) that can take on the order of hours to days. Furthermore, analysis by Raman spectroscopy does not require chemical labeling and only minimal sample preparation (i.e., applying μL amounts of culture to an aluminum surface). This is not true for traditional methods of analysis, which require steps such as cell fractionation, extraction, and derivatization. Several challenges face Raman spectroscopy of microbial cells, one of which is data normalization and the absence of good internal standards. This has been addressed in the current research by looking at Raman band ratios (i.e., fatty acids to nucleic acids). Generating absolute values of fatty acids production remains a challenge with Raman spectroscopy, but this method can quickly screen candidates to identify those that may be analyzed in detail by traditional analytical methods. Here, Raman spectroscopy was used to analyze the phenotypes of more 1-butanol tolerant cultures containing each of 9 enriched DNA fragments. In the end, 9 different phenotypes were observed, each being significantly different from the pCTL plasmid control strain. Some commonalities in phenotypes were observed such as: (i) increased unsaturated fatty acids composition, (ii) decreased cyclopropane fatty acids composition, (iii) increased membrane rigidity, and (iv) increased ratio of total fatty acids to nucleic acids. Phenotypes differed whether they increased/decreased saturated fatty acids composition and total fatty acids to phenylalanine (total protein) ratio. Chemometric fingerprinting of Raman spectra revealed that no two phenotypes were the same. This information is valuable as the mechanisms of these enriched DNA fragments are sought and researchers explore strategies to rationally engineer fatty acids production and 1-butanol tolerance.

ACKNOWLEDGEMENTS

This research was funded in part by USDA CSREES AFRI (Award: 2010-65504-20346) and NSF MCB (Award: NSF1243988). BF also received funding from the Institute for Critical Technology and Applied Science (ICTAS) at Virginia Tech. BF and RW performed the genomic library preparation, enrichment, and outgrowth. TZ performed Raman spectroscopy analysis. BF, TZ, and RS designed experiments and prepared the manuscript. All authors have read and approved the final version. We also acknowledge contributions Jennifer Jenrette, Robert Settlage, and Saikumar Karyala at the Virginia Bioinformatics Institute for the genomic sequencing data and for the assistance of Lauren Grisso, Jessica Symons, Michelle Halsted, Parker Lee, and Nina Wilson.

REFERENCES

[1] Jones, D. T., Woods, D. R., Acetone-butanol fermentation revisited. *Microbiological reviews* 1986, *50*, 484-524.

[2] Papoutsakis, E. T., Engineering solventogenic clostridia. *Current Opinion in Biotechnology* 2008, *19*, 420-429.

[3] Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., *et al.*, Metabolic engineering of *Escherichia coli* for 1-butanol production. *Metabolic Engineering* 2008, *10*, 305-311.

[4] Lütke-Eversloh, T., Bahl, H., Metabolic engineering of *Clostridium acetobutylicum*: recent advances to improve butanol production. *Current Opinion in Biotechnology* 2011, *22*, 634-647.

[5] Lee, J. Y., Jang, Y. S., Lee, J., Papoutsakis, E. T., Lee, S. Y., Metabolic engineering of Clostridium acetobutylicum M5 for highly selective butanol production. *Biotechnology Journal* 2009, *4*, 1432-1440.

[6] Malinowski, J. J., Two-phase partitioning bioreactors in fermentation technology. *Biotechnology Advances* 2001, *19*, 525-538.

[7] Oudshoorn, A., van der Wielen, L. A. M., Straathof, A. J. J., Assessment of Options for Selective 1-Butanol Recovery from Aqueous Solution. *Industrial & Engineering Chemistry Research* 2009, *48*, 7325-7336.

[8] Lee, S. Y., Park, J. H., Jang, S. H., Nielsen, L. K., *et al.*, Fermentative butanol production by clostridia. *Biotechnology and Bioengineering* 2008, *101*, 209-228.

[9] Nair, R. V., Green, E. M., Watson, D. E., Bennett, G. N., Papoutsakis, E. T., Regulation of the sol Locus Genes for Butanol and Acetone Formation in Clostridium acetobutylicumATCC 824 by a Putative Transcriptional Repressor. *Journal of Bacteriology* 1999, *181*, 319-330.

[10] Tomas, C. A., Welker, N. E., Papoutsakis, E. T., Overexpression of groESL in *Clostridium acetobutylicum* Results in Increased Solvent Production and Tolerance, Prolonged Metabolism, and Changes in the Cell's Transcriptional Program. *Appl. Environ. Microbiol.* 2003, *69*, 4951-4965.

[11] Borden, J. R., Papoutsakis, E. T., Dynamics of genomic-library enrichment and identification of solvent tolerance genes for Clostridium acetobutylicum. *Appl Environ Microbiol* 2007, *73*, 3061-3068.

[12] Alsaker, K. V., Spitzer, T. R., Papoutsakis, E. T., Transcriptional Analysis of spo0A Overexpression in *Clostridium acetobutylicum* and Its Effect on the Cell's Response to Butanol Stress. *J. Bacteriol.* 2004, *186*, 1959-1971.

[13] Pich, A., Narberhaus, F., Bahl, H., Induction of heat shock proteins during initiation of solvent formation inClostridium acetobutylicum. *Appl Microbiol Biotechnol* 1990, *33*, 697-704.

[14] Hermann, M., Fayolle, F., Marchal, R., Podvin, L., *et al.*, Isolation and characterization of butanol-resistant mutants of *Clostridium acetobutylicum*. *Appl. Environ. Microbiol.* 1985, *50*, 1238-1243.

[15] Terracciano, J. S., Rapaport, E., Kashket, E. R., Stress and Growth Phase Associated Proteins of *Clostridium acetobutylicum*. *Appl. Environ. Microbiol.* 1988, *54*, 1989-1995.

[16] Zhang, H., Chong, H., Ching, C., Song, H., Jiang, R., Engineering global transcription factor cyclic AMP receptor protein of Escherichia coli for improved 1-butanol tolerance. *Appl Microbiol Biotechnol* 2012, *94*, 1107-1117.

[17] Chin, W.-C., Lin, K.-H., Chang, J.-J., Huang, C.-C., Improvement of n-butanol tolerance in *Escherichia coli* by membrane-targeted tilapia metallothionein. *Biotechnology for Biofuels* 2013, *6*, 130.

[18] Liu, S., Qureshi, N., How microbes tolerate ethanol and butanol. *New Biotechnology* 2009, *26*, 117-121.

[19] Woodruff, L. B. A., Pandhal, J., Ow, S. Y., Karimpour-Fard, A., *et al.*, Genome-scale identification and characterization of ethanol tolerance genes in *Escherichia coli*. *Metabolic Engineering* 2013, *15*, 124-133.

[20] Atsumi, S., Wu, T. Y., Machado, I. M., Huang, W. C., *et al.*, Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. *Mol Syst Biol* 2010, *6*, 449.

[21] Vollherbst-Schneck, K., Sands, J., Montenecourt, B., Effect of butanol on lipid composition and fluidity of *Clostridium acetobutylicum* ATCC 824. *Applied and Environmental Microbiology* 1984, *47*, 193-194.

[22] Bowles, L. K., Ellefson, W. L., Effects of butanol on Clostridium acetobutylicum. *Applied and Environmental Microbiology* 1985, *50*, 1165-1170.

[23] Van Der Westhuizen, A., Jones, D. T., Woods, D. R., Autolytic Activity and Butanol Tolerance of *Clostridium acetobutylicum*. *Applied and Environmental Microbiology* 1982, *44*, 1277-1281.

[24] Ounine, K., Petitdemange, H., Raval, G., Gay, R., Regulation and butanol inhibition of D-xylose and D-glucose uptake in Clostridium acetobutylicum. *Applied and Environmental Microbiology* 1985, *49*, 874-878.

[25] Kanno, M., Katayama, T., Tamaki, H., Mitani, Y., *et al.*, Isolation of butanol and isobutanol tolerant bacteria and physiological characterization of their butanol tolerance. *Applied and Environmental Microbiology* 2013.

[26] Borden, J. R., Jones, S. W., Indurthi, D., Chen, Y., Papoutsakis, E. T., A genomic-library based discovery of a novel, possibly synthetic, acid-tolerance mechanism in Clostridium

acetobutylicum involving non-coding RNAs and ribosomal RNA processing. *Metab Eng* 2010, *12*, 268-281.

[27] Lynch, M. D., Warnecke, T., Gill, R. T., SCALEs: multiscale analysis of library enrichment. *Nat Methods* 2007, *4*, 87-93.

[28] Reyes, L. H., Almario, M. P., Kao, K. C., Genomic library screens for genes involved in n-butanol tolerance in Escherichia coli. *PloS one* 2011, *6*, e17678.

[29] González-Ramos, D., van den Broek, M., van Maris, A. J., Pronk, J. T., Daran, J.-M. G., Genome-scale analyses of butanol tolerance in *Saccharomyces cerevisiae* reveal an essential role of protein degradation. *Biotechnology for Biofuels* 2013, *6*, 48.

[30] Sambrook, J., Russell, D. W., *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press 2001.

[31] Leland, C., Peter, K., Genomic DNA Libraries, *Handbook of Molecular and Cellular Methods in Biology and Medicine, Second Edition*, CRC Press 2003.

[32] Kittler, R., Stoneking, M., Kayser, M., A Whole Genome Amplification Method to Generate Long Fragments from Low Quantities of Genomic DNA. *Analytical Biochemistry* 2002, *300*, 237-244.

[33] Pinard, R., de Winter, A., Sarkis, G., Gerstein, M., *et al.*, Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006, *7*, 216.

[34] Uda, A., Tanabayashi, K., Fujita, O., Hotta, A., *et al.*, Comparison of whole genome amplification methods for detecting pathogenic bacterial genomic DNA using microarray. *Japanese journal of infectious diseases* 2007, *60*, 355.

[35] Zu, T. N., Athamneh, A. I., Wallace, R. S., Collakova, E., Senger, R. S., Near real-time analysis of the phenotypic responses of Escherichia coli to 1-butanol exposure using Raman spectroscopy. *J Bacteriol* 2014.

[36] Athamneh, A. I., Alajlouni, R. A., Wallace, R. S., Seleem, M. N., Senger, R. S., Phenotypic profiling of antibiotic response signatures in Escherichia coli using Raman spectroscopy. *Antimicrobial agents and chemotherapy* 2014, *58*, 1302-1314.

[37] Crow, P., Stone, N., Kendall, C. A., Uff, J. S., *et al.*, The use of Raman spectroscopy to identify and grade prostatic adenocarcinoma in vitro. *British journal of cancer* 2003, *89*, 106-108.

[38] Mahadevan-Jansen, A., Mitchell, M. F., Ramanujam, N., Malpica, A., *et al.*, Near-infrared Raman spectroscopy for in vitro detection of cervical precancers. *Photochemistry and photobiology* 1998, *68*, 123-132.

[39] Dutta, R. K., Sharma, P. K., Pandey, A. C., Surface enhanced Raman spectra of Escherichia Coli cell using ZnO nanoparticles. *Dig. J. Nanomater. Biostruct* 2009, *4*, 83-87.

[40] She, C., Dinh, N., Tu, A. T., Laser raman scattering of glucosamine *N*-acetylglucosamine, and glucuronic acid. *Biochimica et Biophysica Acta (BBA)-General Subjects* 1974, *372*, 345-357.

[41] Minty, J. J., Lesnefsky, A. A., Lin, F., Chen, Y., *et al.*, Evolution combined with genomic study elucidates genetic bases of isobutanol tolerance in Escherichia coli. *Microbial cell factories* 2011, *10*, 18.

[42] Papoutsakis, E. T., Engineering solventogenic clostridia. *Current opinion in biotechnology* 2008, *19*, 420-429.

[43] Dunlop, M. J., Dossani, Z. Y., Szmidt, H. L., Chu, H. C., *et al.*, Engineering microbial biofuel tolerance and export using efflux pumps. *Mol Syst Biol* 2011, *7*, 487.

[44] Read, D. S., Woodcock, D. J., Strachan, N. J., Forbes, K. J., *et al.*, Evidence for phenotypic plasticity among multihost Campylobacter jejuni and C. coli lineages, obtained using ribosomal multilocus sequence typing and Raman spectroscopy. *Appl Environ Microbiol* 2013, *79*, 965-973.

[45] Harris, L. M., Welker, N. E., Papoutsakis, E. T., Northern, Morphological, and Fermentation Analysis of spo0A Inactivation and Overexpression in Clostridium acetobutylicum ATCC 824. *J. Bacteriol.* 2002, *184*, 3586-3597.

[46] Bimboim, H. C., Doly, J., A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* 1979, *7*, 1513-1523.

[47] Tang, X., Nakata, Y., Li, H.-O., Zhang, M., *et al.*, The optimization of preparations of competent cells for transformation of E. coli. *Nucleic Acids Research* 1994, *22*, 2857.

[48] Gordon, A., Hannon, G., Fastx-toolkit: FASTQ/A short-reads pre-processing tools. *(unpublished) http://hannonlab.cshl.edu/fastx_toolkit* 2010.

[49] Bolger, A. M., Lohse, M., Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, btu170.

[50] Langmead, B., Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012, *9*, 357-359.

[51] Li, H., Handsaker, B., Wysoker, A., Fennell, T., *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, *25*, 2078-2079.

[52] García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., *et al.*, Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012, *28*, 2678-2679.

[53] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., *et al.*, The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bmc Bioinformatics* 2008, *9*, 386.

[54] Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., Kschischo, M., grofit: fitting biological growth curves with R. *Journal of Statistical Software* 2010, *33*, 1-21.

[55] Nicolaou, S. A., Gaida, S. M., Papoutsakis, E. T., Coexisting/Coexpressing Genomic Libraries (CoGeL) identify interactions among distantly located genetic loci for developing complex microbial phenotypes. *Nucleic Acids Research* 2011, *39*, e152-e152.

[56] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J Mol Biol* 1990, *215*, 403-410.

[57] Movasaghi, Z., Rehman, S., Rehman, I. U., Raman spectroscopy of biological tissues. *Appl Spectrosc Rev* 2007, *42*, 493-541.

[58] Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., *et al.*, The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *Journal of Bacteriology* 2008, *190*, 2597-2606.

[59] Nolling, J., Breton G Fau - Omelchenko, M. V., Omelchenko Mv Fau - Makarova, K. S., Makarova Ks Fau - Zeng, Q., *et al.*, Genome sequence and comparative analysis of the solvent-producing bacterium. *J Bacteriol* 2001, *183*, 4823-4838.

[60] Borden, J. R., Papoutsakis, E. T., Dynamics of Genomic-Library Enrichment and Identification of Solvent Tolerance Genes for Clostridium acetobutylicum. *Applied and Environmental Microbiology* 2007, *73*, 3061-3068.

[61] Fontaine, F., Stewart, E. J., Lindner, A. B., Taddei, F., Mutations in two global regulators lower individual mortality in *Escherichia coli*. *Molecular Microbiology* 2008, *67*, 2-14.

[62] Peng, W., Takabayashi, H., Ikawa, K., Whole genome amplification from single cells in preimplantation genetic diagnosis and prenatal diagnosis. *European Journal of Obstetrics & Gynecology and Reproductive Biology* 2007, *131*, 13-20.

[63] Martino, A., Rhodes, M. E., Biddle, J. F., Brandt, L. D., *et al.*, Novel degenerate PCR method for whole genome amplification applied to Peru Margin (ODP Leg 201) subsurface samples. *Frontiers in Microbiology* 2012, *3*.

[64] Reeder, T., Schleif, R., Mapping, sequence, and apparent lack of function of araJ, a gene of the Escherichia coli arabinose regulon. *Journal of Bacteriology* 1991, *173*, 7765-7771.

[65] Fritz, G., Megerle, J. A., Westermayer, S. A., Brick, D., *et al.*, Single Cell Kinetics of Phenotypic Switching in the Arabinose Utilization System of E. coli. *PloS one* 2014, *9*.

[66] H, J., V, B., CH, L., D, V., *et al.*, Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol* 2009, *394*, 644-652.

[67] RM, F., TM, R., Saier MH, J., The mannitol repressor (MtlR) of Escherichia coli. *J Bacteriol* 1994, *176*, 840-847.

[68] Denison, M. S., Whitlock, J. P., Xenobiotic-inducible Transcription of Cytochrome P450 Genes. *Journal of Biological Chemistry* 1995, *270*, 18175-18178.

[69] Fontaine, F., Stewart, E. J., Lindner, A. B., Taddei, F., Mutations in two global regulators lower individual mortality in Escherichia coli. *Molecular microbiology* 2008, *67*, 2-14.

[70] Asakura, Y., Kobayashi, I., From damaged genome to cell surface: transcriptome changes during bacterial cell death triggered by loss of a restriction–modification gene complex. *Nucleic Acids Research* 2009, *37*, 3021-3031.

Chapter 4

A genomic library to alter the metabolic profile of *Clostridium cellulolyticum* ATCC 35319

ABSTRACT

Clostridium cellulolyticum possess the ability to ferment cellulosic substrates to generate high value chemicals and commodities from abundant feed-stocks. The growth and alcohol production rates of the wild-type organism are not suitable for industrial utilization without metabolic engineering. A genomic expression library was produced by a degenerate oligonucleotide primed (DOP) PCR amplification method to demonstrate its effectiveness in high-throughput expression studies. A partial coverage library originating from the *C. cellulolyticum* genome and an environmental soil sample from the Virginia Tech campus were generated from this method, sequenced using the Illumina® MiSeq, and expressed in *C. cellulolyticum* subjected to a competition assay of serial dilutions in a cellobiose based media to isolate library fragments imparting a fitness advantage. The resulting isolation identified a conserved phage structural protein fragment which enabled the organism to produce 370% more ethanol than the control strain or fragments enriched from the metagenomic sample by decreasing production of lactate and resulted in 350% greater maximum accumulation of pyruvate. This experiment demonstrated the advantages of using high-throughput library enrichment of uncharacterized genomes to identify un-predicted and novel targets for genetic engineering outside the sample space of targeted engineering approaches.

INTRODUCTION

The role of lignocellulosic biofuels

Early industrial bioprocessing involved the production of simple alcohols from simple sugars and molasses. Market forces have since necessitated technological innovations in order to produce large quantities of more valuable chemicals from cheaper and recalcitrant feed stocks. Lignocellulose is a heterogeneous polymer of semi-regular hexose and pentose sugar molecules interspersed with large, irregular phenylpropynol complexes. The combined quantity of hexose sugars approaches 10^{14} kg in the natural ecosystem [1]. The abundance of raw lignocellulose material is complicated by the structural stability of the intertwined network of cellulose, hemicellulose, and lignin. Despite the difficulties, an economic outlook on the future of bioprocessing has indicated the industry *must* take advantage of the sheer volume of lignocellulosic feed stocks to remain cost competitive with petroleum [2]. Lignocellulose has been used as a feed stock for many value-added products, including hydrogen (H_2) [3]; antibiotics [4]; isobutanol [5]; acetone, butanol, and ethanol (ABE) [6]; and 12 of the critical chemical targets identified by the US Department of Energy [7]. The effective incorporation of lignocellulose into traditional bioprocessing fermentations is predicted to hinge on the advancement of consolidated bioprocessing, which is an engineering approach to simultaneously digest lignocellulose and ferment its liberated sugars into biofuels and valuable chemicals [8-10].

Processing of lignocellulose has been performed chemically for over a century, dating back to early Kraft pulping and paper manufacturing. Traditionally, chemical pretreatment separates the polymers of lignocellulose and opens the cellulose chains to enzymatic and microbial digestion [11, 12]. Following pretreatment, hydrolysis of the 5 and 6 carbon polymer chains takes place, and the liberated sugars are fermented into alcohol. A consolidated

bioprocessing operation combines the hydrolysis and fermentation steps, relying on cellulolytic organisms to perform the hydrolysis and fermentation in a single stage [2]. The Gram positive obligate anaerobe *Clostridium cellulolyticum* ATCC 35319 is of interest to consolidated bioprocessing due to its ability to breakdown and consume crystalline cellulose under mesophilic conditions [13-15]. It breaks-down cellulose through the use of a cellulosome, and ethanol is a major fermentation product. Additionally, *C. cellulolyticum* has been shown to express specific enzymes for the digestion of 5-carbon sugars [16]. However, it is characterized by a slow growth rate, and it secretes the valuable central carbon metabolite pyruvate as a byproduct. Unlike several other clostridia, it has proven amenable to genetic engineering [17], and the focus of this research is to alleviate metabolic bottlenecks and improve acid and alcohol tolerance through the use of combinatorial metabolic engineering with genomic DNA libraries.

Cellulolytic clostridia and improvements through metabolic engineering

The mechanism of cellulose digestion comes from the extracellular cellulosome complex: an interchangeable combination of proteins that work in synergy to facilitate cellulose breakdown. The *C. cellulolyticum* cellulosome is composed of a scaffold protein, CipC, with up to eight "modules" for attachment of different enzymes. Each enzyme module contains a unique dockerin domain which adheres to the scaffold through interaction with a specific, complimentary cohesion domain [18]. In addition to cellulose binding domains, thirteen separate catalytic domains have been identified, consisting of cellobiohydrolases, endoglucanases, cellulases, a pectinase, a mannanase and a β-xylosidase [19] where the majority of genes lie in a single *cel* DNA operon [20-23]. This complex allows *C. cellulolyticum* to hydrolyze and digest

glucose, xylose, arabinose, galactose, mannose, and rhamnose from the lignocellulose complex [22-24].

A close relative of *C. cellulolyticum*, the thermophile *C. thermocellum*, has a genetically similar cellulosome structure, differing only by an additional calcium requirement for cellulose binding [23]. The dockerin domains between the two species do not bind to the other's cohesion domains and are not interchangeable, however, without amino-acid substitutions in the dockerin protein sequence [25]. Expression of functional and compatible dockerin and cohesion elements in non-host organisms has been performed, opening the door for designer cellulosomes whose composition is tailored to optimal digestion of the substrate [26, 27].

Despite the remarkable machinery for cellulose digestion, *C. cellulolyticum* suffers from an inefficient carbon flow that limits its growth rate and productivity for practical purposes. The generation (doubling) time of wild-type *C. cellulolyticum* ATCC 35319 is approximately 4 hours on soluble cellobiose and 8-13 hours on whole crystalline cellulose [28, 29], while the alcohol production rate has been reported as 1.51 mmol $gDCW^{-1}$ h^{-1} (~7.0 mg ethanol L^{-1} hr^{-1}) in batch [30] and 1.27 mmol $gDCW^{-1}$ hr^{-1} (9.4 mg ethanol L^{-1} h^{-1}) in continuous production [31]. In comparison, the desired characteristics for industrial microbial fermentation includes an alcohol tolerance greater than 40 g/L and ethanol productivity greater than 1 g L^{-1} h^{-1} [32]. Part of the reason for the relatively poor fermentative capabilities of *C. cellulolyticum* is believed to result from adaptation of the cell to growth in low-nutrient environments. During growth, glucose-1-phosphate and glucose-6-phosphate accumulate in *C. cellulolyticum* due to the unregulated uptake of the cellulose disaccharide, cellobiose [33]. Under low carbon flow, the organism optimizes ATP and NADH formation through conversion of pyruvate to acetate and ethanol. However, during carbon excess, ATP/ADP and NADH/NAD+ ratios increase, inhibiting the

glyceraldehyde-3-phosphate dehydrogenase enzyme in glycolysis. *C. cellulolyticum* was found to respond to such stress by the accumulation of intracellular pyruvate (as well as secretion of pyruvate) and extended lactate production at the expense of acetate and ethanol [34]. To combat this shortcoming, targeted metabolic engineering of *C. cellulolyticum* has been used to increase gene expression of pyruvate consuming enzymes pyruvate decarboxylase and alcohol dehydrogenase. The resulting phenotype was characterized by a (i) 150% increase in cellulose consumption rate, (ii) 180% increase in biomass, (iii) 48% reduction in lactate concentrations, and (iv) 2x increase ethanol titers [35]. One of the first trials of gene knock-out in *C. cellulolyticum* resulted in a strain capable of producing 8.5x as much ethanol from cellulose by silencing the genes for both lactate and malate dehydrogenases. These are competing enzymes that divert pyruvate from ethanol. This increase in ethanol occurred at the expense of lactate and acetate production and produced a culture with a slower growth rate [36]. From an industrial perspective, wild-type *C. cellulolyticum* has both a poor growth rate and alcohol productivity. However, its value is with the utility and completeness of its cellulolytic complex. This presents researchers with two options: (i) derive and apply metabolic engineering strategies to make *C. cellulolyticum* a more robust organism for lignocellulose fermentation or (ii) utilize the cellulosome complex of *C. cellulolyticum* in other organisms that lack lignocellulolytic capabilities but are capable of industrial fermentation. The machinery of the *Clostridium* cellulosome has been cloned and expressed in other host systems, including *Clostridium acetobutylicum* [37], *Escherichia coli* [38], and *Saccharomyces cerevisiae* [39]. Some modified yeast constructs have achieved ethanol production rates on cellulose comparable to *C. cellulolyticum* [8, 27]. However, the focus of the research presented here is to increase the

growth rate of *C. cellulolyticum* using a combinatorial metabolic engineering approach with genomic DNA libraries.

Combinatorial approach for rapid screening of multiple genetic variants

Targeted approaches have been useful in increasing growth and solvent production rates of bacteria when the metabolic networks are well characterized. When the network is not well understood or when multiple genes control a phenotype synergistically, a combinatorial metabolic engineering stategy can be quite effective. This strategy involves creating a large mutant pool and enriching for the traits of interest, usually in the form of increased growth rate or tolerance to a growth inhibitor. Mutants developing solvent tolerance survive when challenged at increasing solvent concentrations, out-competing non-tolerant strains [40, 41]. When no stressors are present, cultures select for traits that produce faster growth rates or greater utilization of substrate in order to out-compete neighboring cells for available nutrients. Combinatorial DNA libraries have been generated with chemical mutagens [42], biological mutagens [43], or overexpression of genomic DNA fragments [2, 44, 45]. For example, Blouzard et al. [46] used the phage Tn1545 transposon in *C. cellulolyticum* allowing for generation of cellulose-fermenting mutants by genomic rearrangement inside the living cell. No selective growth assay has yet been performed on the mutants. Gill et al. [45] developed the multi-Scalar Analysis of Library Enrichments (SCALEs) method, which uses combinatorial libraries and DNA microarrays to identify library fragments that become enriched (i.e., contribute to cellular fitness) during growth inhibition assays. By adapting the culture to increasing inhibitor concentrations, researchers have used the SCALEs approach to determine cellular response to ethanol [47], 3-hydroproponic acid [48], aspartic acid [49], 1-napthol [50], and antimicrobials

[45]. Borden et al. [51, 52] prepared and fractionalized genomic DNA from *C. acetobutylicum* to be repackaged into expression vectors. This technique was used to rapidly screen for substrains containing overexpressed genes or DNA fragments. A 16S rRNA promoter region was identified that conferred up at an 81% increase in tolerance to 1-butanol.

Metabolic engineering of *C. cellulolyticum* with genomic DNA libraries

The use of combinatorial genomic DNA libraries has produced success in resolving auxotrophy and increasing solvent tolerance [44, 53]. However, the traditional methods of DNA overexpression library construction was found to be inefficient in our experience. In response, we have developed novel methodology to generate genomic DNA libraries from multiple sources using degenerate oligonucleotide primed PCR (DOP-PCR) amplification of nanogram quantities of genomic material with great repeatability [54]. This novel method was used to generate multiple genomic DNA libraries for *C. cellulolyticum* to search for regulatory or coding DNA sequences to resolve metabolic bottlenecks, increase the growth rate of the organism on cellobiose, and improve acid and alcohol tolerance. Two DNA libraries were produced in this research: (i) from *C. cellulolyticum* genomic DNA and (ii) from an environmental soil metagenome.

MATERIALS AND METHODS

Strains

High efficiency *E. coli* NEB 10-beta competent cells were supplied by New England Biolabs (Ipswich, MA) for library construction. *Clostridium cellulolyticum* ATCC 35319 was

obtained from the American Type Culture Collection. Plasmids and strains are listed in Table 4-1.

Table 4-1. List of cell strains, plasmids, and DNA oligonucleotides used in the study.

Strain	Genotype	Reference
Clostridium cellulolyticum (H10) ATCC 35319	Wild-type	(Petitdemange, Caillet et al. 1984)
Escherichia coli NEB 10-beta	Δ*(ara-leu) 7697 araD139 fhuA* Δ*lacX74 galK16 galE15 e14-* ϕ*80*d*lacZ*Δ*M15 recA1 relA1 endA1 nupG rpsL* (Str^R) *rph spoT1* Δ*(mrr-hsdRMS-mcrBC)*	New England Biolabs
Escherichia coli NEB 5-alpha	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	New England Biolabs
Escherichia coli Ccdb survival 2 T1^{TM}	F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*Δ139 Δ(*ara-leu*)7697*gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G *fhu*A::*IS2*	Life Technologies
Plasmids		
Psos95	Em^r, Amp^r, *thi* promoter, ColE1 ORI, pIM13 ORI	(Girbal, Mortier-Barrière et al. 2003)
PsosLink	Em^r, Amp^r, *thi* promoter	This paper
PsosGate	Em^r, Amp^r, Cm^r, *ccdb* Gateway® Cassette	This paper
pCR8/GW/TOPO TA	Spn^r	Life Technologies
Primers		
9NATG	TAGACAATGGGGCATNNNNNNNNNATG	
MCS1	TTGGGATCCCTAACTAACTAGACACTAAGTC GATATCGACATAGTGTCTAGATAGATAGGGC GCCTTG	
pSOS_libF	TACGGGGTAACAGATAAACCATT	
pSOS_libR	GATAGATAGGGCGCCACTTA	
Gateway_F	ATCACAAGTTTGTACAAAAAAGC	
Gateway_R	ATCACCACTTTGTACAAGAAAGC	

Media and cultivation

E. coli K-12 NEB 10-beta were grown in Luria-Burtani (LB) medium at 37°C and shaken at 225 rpm. Selective media contained 100 mg/L ampicillin where appropriate. *C. cellulolyticum* was maintained on GS-2 media [55] with cellobiose (Sigma-Aldrich; St. Louis, MO) (5g/L) as the carbon source. Cellobiose sugar solution and mixture of Mg^{2+}, Ca^{2+}, and Fe^{3+} salts were autoclaved separately at 10x concentration. After addition of remaining materials except L-cysteine, pH was adjusted to 7.2 and the solution was heated until resazurin was activated, as indicated by light pink color, and autoclaved. L-cysteine solution (20% w/v) was autoclaved separately and 5 mL was added per L of media in Bactron anaerobic chamber (Shel Lab; Cornelius, OR). The anaerobic chamber was maintained at an atmosphere of $N_2/CO_2/H_2$ (90/5/5%). Culture media was allowed to de-gas for 24 h before use. Cells were cultivated at 34°C. Erythromycin was added as the selective reagent from a 1000x concentrated ethanol solution to 10 mg/L and agar was added at 1.5% (m/v) where appropriate.

DNA manipulation

All molecular biology enzymes and reagents were obtained from New England Biolabs (Ipswitch, MA) unless noted otherwise. All PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA). In addition, details of all strains, plasmids, and primer sequences are given in Table 4-1. Plasmid pSOSlink was created from plasmid pSOS95 [56] by digestion with BamHI and EcoRI restriction enzymes, followed by ligation with T4 DNA ligase to similarly digested multiple cloning cassette, MCS1. Plasmids pSOSGate was constructed by ligating the blunt ended T4 polynucleotide kinase phosphorylated Gateway Cloning Cassette (Life Technologies; Grand Island, NY) into EcoRV digested and antarctic phosphatase treated

pSOSlink. This was then transformed into *ccdb* Survival 2 T1™ cells (Life Technologies) selected for on 50 mg/L chloramphenicol media.

Clostridium cellulolyticum genomic DNA was harvested based on a published protocol [57] as follows. An overnight culture of 50 mL of *C. cellulolyticum* in CM3 media was harvested at 5000 rpm and 4°C for 10 min in an Eppendorf 5810R centrifuge. Cells were washed twice with 10 mL of solution containing EDTA (0.1 M), Tris-HCL (0.05 M), and KCL (0.5 M). The washed pellet was then digested in 4 mL of solution containing NaCl (0.1 M), EDTA (0.05 M), egg-white lysozyme (Sigma-Aldrich) (10 mg/μL), and RNase (Qiagen; Valencia, CA) (40 μg/mL) for 30 min at 37°C. Next, 300 μL of 20% SDS was added and the DNA was extracted with Tris-buffered phenol/chloroform/isoamyl alcohol (25/24/1) twice followed by a single extraction with chloroform. The DNA was precipitated with 1 vol isopropanol and 0.2 vol of 3 M sodium acetate. The washed and dried DNA was dissolved in TE buffer overnight.

Whole genome amplification by DOP-PCR

DOP-PCR DNA library construction was performed using oligonucleotide primers containing degenerate (equal probability of A/T/G/C bases) regions to allow for partial and mismatched annealing with random segments of genomic DNA. The primer 9NATG was used in a thermocycling reaction with genomic DNA and *Taq* polymerase according to previously published protocols [58, 59] with modifications that have been published elsewhere [54]. Briefly, whole genomic DNA was diluted to 10 ng/μL. DNA concentration was determined by NanoDrop® 2000c (ThermoFisher Scientific; Waltham, MA). Next, 1 μL of DNA was combined in a chilled 50 μL reaction with 5 μL of 10x Thermopol reaction buffer, 1 μL of 9NATG (100 μM) primer, 1.25 μL dNTP (10 mM each), 1 μL 5U/μL *Taq* Polymerase in Type I

ultra-pure water. The reaction was transferred to an Eppendorf gradient MasterCycler PCR under a mixed regime of non-stringent and high temperature cycling: 6 min initial denaturation at 95°C followed by 10 cycles of (i) 1 min denaturation at 95°C, (ii) 3 min annealing at 30°C, (iii) 3 min ramp to 72°C, (iv) 3.5 min elongation at 72°C. This was followed immediately by 25 cycles of: (i) 1 min denaturation at 95°C, (ii) 1 min annealing at 55°C, (iii) 3 min elongation at 72°C. Then, a final elongation step of 10 min at 72°C was carried out. Products were loaded onto an ethidium bromide Tris-Acetate agarose gel and DNA above 1000 bp was excised and purified in a GeneJet spin column (ThermoFisher Scientific) and diluted to 50 ng/μL.

Genomic DNA library construction

Linearized vector pCR8/GW/TOPO TA (Life Technologies) was combined with 0.5 μL of the DNA mix following manufacturer's directions and incubated for 1 h at 25°C before transforming 2μL into 50 μL of Clonetech® Stellar Competent Cells according to manufacturer's directions. Cells were pooled from 20 individual 50 μL transformations to generate 40,000 colonies and grown up in 200 mL of LB with 50 mg/L spectinomycin to OD_{600} of 0.6 before collecting for midipreps (Gerard Biotech; Oxford, OH) and creating frozen stocks (stored at -80°C).

Plasmid DNA in pCR8 was recombined with pSOSGate by combining 150 ng of plasmid with 50 ng of the pooled pCR8 library DNA using the Invitrogen Gateway LR Clonase II kit and incubating overnight at 25°C. Aliquots of 1 μL transformation mix were used to transform ten 50 μL vials of *E. coli* NEB 5-alpha competent cells, and transformants were pooled once again, generating 100,000 mutants in LB media with ampicillin (100 mg/L) selection. Cells were grown to late log phase and DNA was harvested using the alkaline lysis procedure [60].

Metagenomic DNA library construction

Metagenomic Genomic DNA from a stream bank soil sample on the Virginia Tech campus in Blacksburg, VA was collected as described in Chapter 2. The extracted genomic DNA was amplified by DOP-PCR as described above. The resulting DNA fragments were cloned into pCR8/GW/TOPO TA (also as described above) to generate 80,000 mutants. The plasmid library was recombined with plasmid pSOSGate using the Gateway LR Clonease II reaction mix and transformed into *E. coli* NEB 5-alpha to generate 120,000 mutants.

Electro-transformation

A total of 25 μg of plasmid DNA was purified using PCR purification columns (ThermoFisher Scientific) and methylated with 20 U MspI methyltransferase in a 200 uL reaction volume with 32 mM S-adenosylmethionine and 20 μL MspI reaction buffer overnight at 37°C. The reaction was purified and concentrated in a PCR purification column to 200 ng/μL before electroporation into *Clostridium cellulolyticum* using the methods of Cui et al [61]. Following incubation with 10 mg/mL glycine, 50 mL of *C. cellulolyticum* culture (OD_{600} = 0.60) was washed twice and suspended in 1.5 mL of 270 mM sucrose and 5 mM $NaPO_4$ (pH 7.4). A volume of 500 μL of cells were combined with 1-2 μg of DNA and electroporated in a BioRad Gene Pulser at 1000 V cm^{-1}, 25 μF, and 750 Ω with a time constant of 8.5-9.5 ms. Cells were immediately suspended in GS-2 media with cellobiose and incubated overnight. After outgrowth, representative plates were made on solid GS-2 media with 10 mg/L erythromycin and remaining cells were transferred to GS-2 media with 10 mg/L erythromycin until OD_{600} of 0.4 was reached. Then frozen stocks were made. All transformed *C. cellulolyticum* were pooled for the outgrowth

library. Pooling of 100 electro-transformed samples of the methylated library fragments in pSOSgate indicated approximately 2,000 mutants existed within the culture. Electroporation of the environmental DNA was performed to generate an additional 2,000 colonies as measured by plating on GS-2 agar with 20 mg/L erythromycin.

Outgrowth and enrichment

The *C. cellulolyticum* and metagenomic DNA library frozen stocks were grown in separate overnight cultures. Both cultures, along with a separate control population consisting solely of pSOSlink were grown to mid-log phase (OD 0.6) and diluted 1:1000 into fresh GS-2 media with antibiotics in an anaerobic environmental chamber at 34°C in static flasks. A fresh dilution was made every 24 h and both cultures were plated out after 5 re-inoculations. Individual colonies were screened and sequenced. Four mutants were chosen for further testing. Two are fragments from the *C. cellulolyticum* genome (H10_phage, H10_Bk) and two are from the metagenomic outgrowth (ENV_M1, ENV_F1). The insert-bearing plasmids were isolated and retransformed into unchallenged *C. cellulolyticum* cultures.

Fresh cultures of individual mutants and the control strain bearing pSOSlink were inoculated in triplicate into 50 mL of GS-2 erythromycin media to a normalized OD_{600} of 0.025. Samples were taken every 12 hours for three days. Samples were measured for OD_{600} and metabolites, as discussed below.

Metabolite analysis

Concentrations of extracellular ethanol, cellobiose, lactic acid, and acetate were measured with a Bio-Rad Aminex® HPX-87H 300x7.8mm column on a Shimadzu HPLC system with an

RID-10A refractive index detector using an isocratic mobile phase of 5 mM sulfuric acid maintained at a constant flow rate of 0.5 mL/min. Whole fermentation samples were filtered through a 0.2 μm filter prior to injection. The sample injection volume was 15 μL and the retention times and concentrations of the peaks were determined with pure standard solutions and calibration curves.

DNA library sequencing and analysis

Next-generation sequencing was performed by the Virginia Bioinformatics Institute using the Illumina MiSeq platform. DNA libraries were prepared for sequencing using the Apollo 324 with IntegenX reagents from DOP amplified, 1-5 kb size selected DNA to generate 350 bp fragments read in 150 bp paired-end mode with a 5% phix spike. Trimming of the primed reads harboring 5' primer sequences was performed with the FASTX toolkit with the following parameters: "Mismatches=2" and "Frameshift=1." Alignment of sequencing reads to the native genome was performed with Bowtie2 against the *C. cellulolyticum* genome (NCBI CP001348.1) and analyzed further with Qualimap v2.0 [64].

Sequence reads from the metagenomic soil DNA was parsed to remove trailing 5' barcode and adapter regions from the 9NATG primer and was uploaded to the MG-RAST webserver [65]. Paired reads were combined and filtered for low quality but not duplicates or contaminating DNA in order to gauge sample bias from the amplified library. The resulting analysis uses BLASTX and BLASTN to compare the uploaded data to multiple nucleotide and protein databases. Results of the analysis are publically available under accession number 4583655.3 on MG-RAST.

RESULTS

DNA library construction by DOP-PCR

DOP-PCR amplification of genomic DNA produced a pool of library fragments that slightly under-represented the reference *C. cellulolyticum* genome. Next-generation sequencing of the initial DNA library indicated 76% base-by-base coverage of the reference genome at 1x coverage and 26% of the genome represented at 5x coverage. These results are shown in Figure 4-.1. The DNA library had an aligned GC content of 39.9%, compared to 37% for the *C. cellulolyticum* reference genome. The metagenomic DNA library fragments were matched to targets on the MG-RAST server with a maximum e-value cutoff of 10^{-5} and minimum identity of 60%. The resulting metagenome had a GC content of 53±10%, represented 4,773 organisms, and contained 569,220 identifiable protein coding regions. These results are summarized further in Table 4-2. The breakdown of sequence targets for the metagenomic DNA library is shown in detail in Figure 7-1 in the Supplementary Appendix.

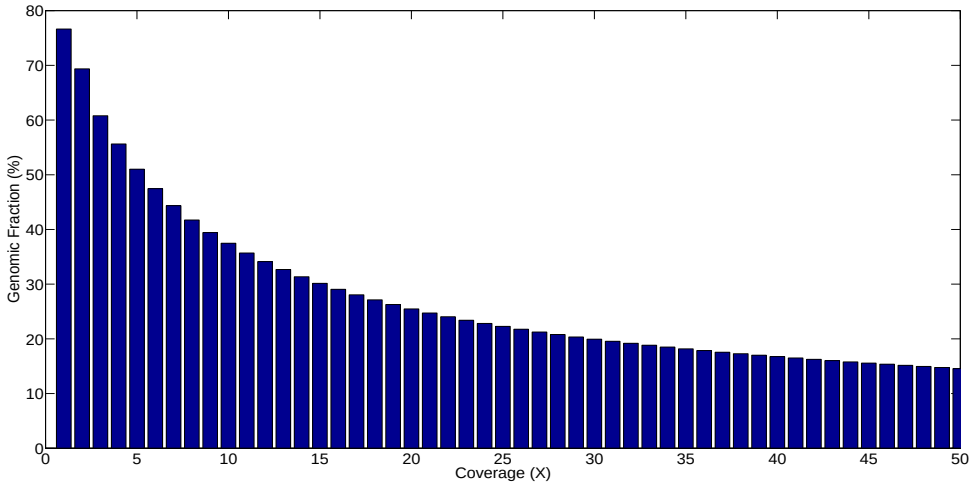


Figure 4-1. Coverage of the DOP-PCR DNA library constructed from *C. cellulolyticum* genomic DNA against the reference genome (NC_011898.1). Coverage refers to the depth or number of times an individual location in the reference is found in the aligned sequencing data. A genomic fraction of 100% at 1X coverage would indicate every base in the reference is represented at least once in the DNA library.

DNA fragments produced from both libraries were ligated into the pCR8/GW/TOPO/TA and produced over 40,000 *E. coli* colonies to generate frozen stocks and 100,000-120,000 clones in the shuttle vector, pSOSGate. The limiting transformation step was electroporation into the *C. cellulolyticum* genome. This procedure achieved less than 2,000 colonies identifiable by plating on solid media. Despite methylation of the plasmid DNA, transformation of over 200 µg of plasmid DNA, and the glycine incubation suggested by Cui et al. (2012) [61], our transformation yields were initially below ten mutant colonies per µg of DNA. We relied on direct liquid inoculation of the transformed cultures without plating to bypass the organism's low plating

efficiency. This has been published elsewhere where it was found that low plating efficiency may actually underrepresent our actual library size by 10^5-fold [66]. However, we remain unsure of this finding and identify this as a potential limitation of the methodology used in this research. However, after five re-inoculations during enrichment, DNA fragments that impart the largest fitness advantage (or least fitness burden) were observed to dominate the culture. Even if the transformed DNA library was limited in size, DNA fragments conferring a growth advantage were present in the DNA library. Enrichment was nearly uniform, as only two unique DNA fragments could be detected in a sample of the *C. cellulolyticum* enriched library (N=10 colonies tested) and two within the metagenomic environmental library (N=18 colonies tested).

Table 4-2. Sequencing results of DOP-PCR amplified genomic and metagenomic DNA used for library construction and enrichment. Protein Coding Features include total annotated genes in reference if available.

DNA Library	Number of Reads	Total Hits (%)	GC Content	Alpha Diversity	Protein Coding Features (%)	Organisms Featured
C. cellulolyticum	3,419,032	3,347,393 (97.3)	39.9%	N/A	3,345 (93.7%)	1
Environmental	2,226,424	1,922,706 (86.4)	53±10%	490.8	569,220 (N/A)	4,773

Table 4-3. List of enriched DNA library fragments after ten subcultures on cellobiose.

Sequence Identifier	Genome Origin[1]	Gene/Protein Similarity
H10_Phage	*C. cellulolyticum* CP001348.1:3379404-3379762	Ccel_2823: Phage minor structural protein
H10_BK	*C. cellulolyticum* CP001348.1:989500-990681	Ccel_0859: Beta-ketoacyl synthase Ccel_0860: Beta-ketoacyl synthase
ENV_F1	*Hydrogenophaga* sp. PBC FR720599.2; *Bordetella bronchiseptica*	*fre:* NADH:Flavin oxioreductase/NADH oxidase; Cytochrome C-type protein
ENV_M1[2]	*Pedosphaera pavula*	*phd:* Prevent-host-death family protein; Hypothetical protein; Ubiquinol-cytochrome C reductase

[1] As determined from BLAST nucleotide search with highest alignment score
[2] No nucleotide alignment possible; transcribed protein matches indicated from highest BLASTx score

Genome alignment of enriched DNA fragments

The isolated DNA fragments from the enrichment experiments are listed in Table 4-3. The DNA fragment H10_BK contains the region spanning two β-ketoacyl synthase genes in a polyketide synthesis cluster in the genome, including most of a polyketide synthase dehydrogenase domain within *Ccel_0859*. DNA fragment H10_phage is a 350 bp section of the N-terminal phage minor structural protein found within the *C. cellulolyticum* genome in a genomic region rich in phage related proteins, possibly inserted during viral recombination events in the organisms evolutionary history [67]. The metagenomic library yielded DNA fragments ENV_F1 and ENV_M1 (Table 4-3). The former codes for a 1.1 kb fragment which BLAST aligns partially with the NADH:Flavin oxioreductase gene of *Hydrogenophaga* sp. PBC and partially with a cytochrome C-type protein of *Bodetella brochioseptica*. The mutant ENV_M1 has no sequence alignment, yet it contains translational reading frames that best match to three separate proteins of *Pdosphaera pavula*: a prevent-host-death family protein, an identified hypothetical protein, and a portion of the ubiquinol-cytochrome C reductase.

Metabolic profiles

The initial fermentation trial with all four mutants and a plasmid control strain (3 biological replicates of each) yielded significantly improved performance from only the strain containing the H10_Phage DNA fragment. While the other fragments (H10_BK, ENV_F1, and ENV_M1) were enriched by the culture, significant differences from the plasmid control strain were not found when the strains were grown independently. From here, the strain containing the H10_phage DNA fragment will be the subject of the analysis. The time-course metabolic profiles of acetate, ethanol, lactate, pyruvate, and cellobiose are shown in Figure 4-2 along with culture

growth. Results of specific production, consumption, and growth rates are also given in Figure 4-2. Of particular interest is the ethanol concentration after 118 hours of fermentation. This was 8.8 mM for the strain expressing the H10_Phage DNA fragment while it was only 2.4 mM for the plasmid control strain. Cellobiose consumption rates were nearly identical. This generated a yield of ethanol produced per cellobiose consumed ($Y_{P/S}$) increase of more than 250% for the engineered strain compared to the control. In addition to similar cellobiose consumption, acetate production was nearly identical for the engineered and control strains. The plots further demonstrate that the H10_Phage mutant accumulated 350% greater levels of pyruvate, peaking at 84 hours commensurate with biomass. The extra carbon appears to have come from lactic acid, which did not accumulate past 11 mM for the mutant. Cellobiose consumption was similar for all strains, depleting completely by 96 hours. The environmental mutant did grow almost 20% lower than the control and reached a lower maximum biomass than then either the mixed culture or the *C. cellulolyticum* based library fragments. The specific production and consumption rates corroborate the ethanol and pyruvate differences as the cultures containing the H10_Phage mutant produce ethanol and pyruvate more rapidly while developing biomass and lactate at a slower than the control (Figure 4-2).

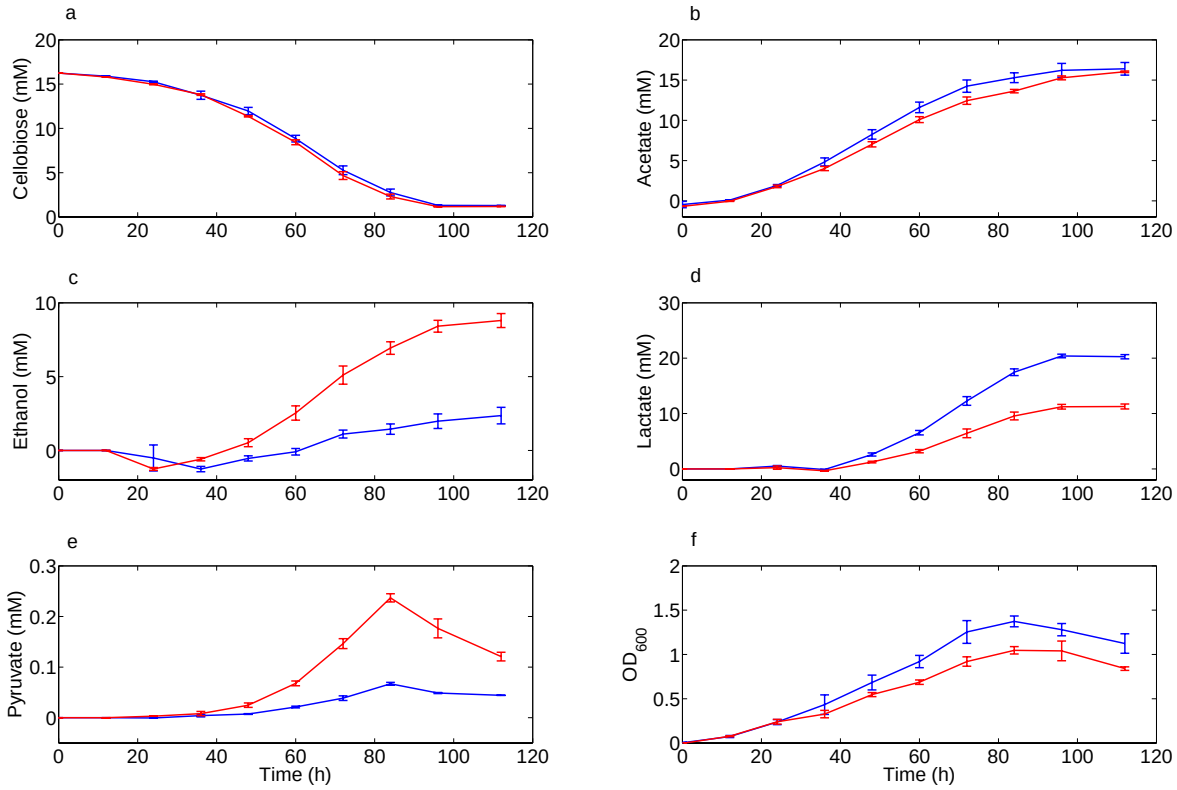


Figure 4-2. Batch fermentation of C. cellulolyticum plasmid control strain (red) and engineered strain containing the H10_Phage DNA fragment (blue) . The following metabolic profiles are shown: Each point represents an average of three biological replicates, and error bars represent one standard deviation.

DISCUSSION

The use of DOP generated DNA fragments to form an enrichment library resulted in isolates without fully functional genes. Though the plasmid pSOSGate contained the strong

thiolase promoter from *C. acetobutylicum* [56] it may have forced the production of RNA from the insert even when no full protein was possible. The RNA may act in a regulatory role or may only represent a metabolic burden to the cell. After enrichment, however, the variety of genomic inserts decreases substantially as mutants with the highest fitness outcompete and outnumber other mutants. Fitness may affect the cell in multiple areas and is not always predictable. Though we expected increased rate of biomass accumulation, we see little positive change after the enrichment, suggesting the "improved" phenotype is acting in other modes, perhaps enabling higher plating efficiency during the final analysis of enrichment samples or export of antibiotic compounds to harm neighboring cells [4]. What we do notice is that the enriched cultures have a variety of effects on the fermentation patterns of ethanol, acetate, lactate, and pyruvate. Hydrogen production is also possible from *C. cellulolyticum* as a means of oxidizing ferredoxin to prevent pyruvate buildup by driving it towards acetyl-CoA [68] and measurement H_2 will prove necessary in future studies.

How the genomic library insert functions within the cell remains to be investigated. H10_BK contains fragments for two β-ketoacyl synthase genes in the fatty acid synthase. This enzyme is responsible for elongation of fatty acids through consumption of acetyl-CoA, possibly affecting acetyl-CoA production from pyruvate and preventing over-accumulation, similar to the proposed effect of amino acid synthesis enzyme overexpression [69].However, the fermentation profile remains similar to the control. Interestingly, the β-ketoacyl synthase also enables antibiotic protection, though primarily against thiolactomycin [70]. The H10_Phage mutant contains a 360bp fragment of the uncharacterized and non-essential phage minor structural protein which is highly conserved in *Clostridium*. The environmental samples both contain a likely cytochrome C type functional group which typically serves in the electron transport chain.

These proteins interact with a large variety of substrates but are typically down-regulated upon exposure to reducing-power-limited substrates such as methanol [71], limiting their utility in a fermentation. Mutant ENV_M1 also contains a fragment of the prevent-host-death (*phd*) protein which in its full form counters the death-on-curing protein responsible for plasmid maintenance [72].

While larger libraries would allow for a more comprehensive screen, cloning and transformation in *C. cellulolyticum* remains a difficult progress. Cui et al. [73] present an improved transformation method using a modified electroporation device to achieve up to 10^4 CFU/µg plasmid DNA. Optimization of transformation of the close relative *C. thermocullum* included many factors not examined here, including controlling temperature of the electroporation cuvette, a 0°C outgrowth period with recovery media, field strengths approaching 25KV/cm, isoniacin pre-treatment, or methylation sensitivity of *E. coli* from which plasmid DNA was isolated [74]. Expression of the organism's native genome could be circumvented entirely using a transposon based mutagenesis system recently developed, where portions of the genome are swapped by a Tn1545 transposable element located on a transformation plasmid. Propagation of the cell and plasmid result in accumulation of mutations greater than the initial number of transformed cells [46]. This method, however, requires extensive sequencing to characterize the mutations and is limited to rearranging existing DNA, not expressing foreign genes.

In summary, we have found yet another application for genomic libraries generated through a degenerate oligonucleotide primed amplification reaction [75, 76]. Though limited in size, the library contained genes coding for uncharacterized phage proteins, oxioreducatses, and cytochrome C type proteins which may have the ability to catalyze multiple reactions within the

cell. Without needing to perform an exhaustive characterization of all library mutants, the enrichment strategy identified 4 mutations that somehow increased cell fitness. One such mutant showed a 3-fold increase ethanol production, decreased acetate production, and earlier accumulation of pyruvate than a control-bearing strain while grown on cellobiose.

REFERENCES

[1] Cox, P. M., Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* 2000, *408*, 184-187.

[2] Lynd, L. R., Zyl, W. H. v., McBride, J. E., Laser, M., Consolidated bioprocessing of cellulosic biomass: an update. *Current Opinion in Biotechnology* 2005, *16*, 577.

[3] Ren, N., Wang, A., Cao, G., Xu, J., Gao, L., Bioconversion of lignocellulosic biomass to hydrogen: Potential and challenges. *Biotechnology Advances* 2009, *27*, 1051-1060.

[4] Lincke, T., Behnken, S., Ishida, K., Roth, M., Hertweck, C., Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium Clostridium cellulolyticum. *Angewandte Chemie* 2010, *122*, 2055-2057.

[5] Higashide, W., Li, Y., Yang, Y., Liao, J. C., Metabolic Engineering of Clostridium cellulolyticum for Production of Isobutanol from Cellulose. *Applied and Environmental Microbiology* 2011, *77*, 2727-2733.

[6] Weber, C., Farwick, A., Benisch, F., Brat, D., *et al.*, Trends and challenges in the microbial production of lignocellulosic bioalcohol fuels. *Appl Microbiol Biotechnol* 2010, *87*, 1303-1315.

[7] Werpy, T. A., Holladay, J. E., White, J. F., Pacific Northwest National Laboratory (PNNL), Richland, WA (US) 2004.

[8] Olson, D. G., McBride, J. E., Joe Shaw, A., Lynd, L. R., Recent progress in consolidated bioprocessing. *Current Opinion in Biotechnology* 2012, *23*, 396-405.

[9] Geddes, C. C., Nieves, I. U., Ingram, L. O., Advances in ethanol production. *Current Opinion in Biotechnology* 2011, *22*, 312-319.

[10] Lynd, L. R., Laser, M. S., Brandsby, D., Dale, B. E., *et al.*, How biotech can transform biofuels. *Nature Biotechnology* 2008, *26*, 169-172.

[11] Biermann, C. J., *Essentials of pulping and papermaking*, Academic Press, San Diego, California 1993.

[12] Galbe, M., Zacchi, G., Pretreatment of lignocellulosic materials for efficient bioethanol production, *Biofuels*, Springer 2007, pp. 41-65.

[13] Petitdemange, E., Caillet, F., Giallo, J., Gaudin, C., Clostridium cellulolyticum sp. nov., a Cellulolytic, Mesophilic: Species from Decayed Grass. *International Journal of Systematic Bacteriology* 1984, *34*, 155-159.

[14] Gehin, A., Gelhaye, E., Petitdemange, H., Adhesion of Clostridium cellulolyticum spores to filter paper. *Journal of Applied Microbiology* 1996, *80*, 187-190.

[15] Desvaux, M., Petitdemange, H., Sporulation of Clostridium cellulolyticum while Grown in Cellulose-Batch and Cellulose-Fed Continuous Cultures on a Mineral-Salt Based Medium. *Microb Ecol* 2002, *43*, 271-279.

[16] Blouzard, J.-C., Coutinho, P. M., Fierobe, H.-P., Henrissat, B., *et al.*, Modulation of cellulosome composition in Clostridium cellulolyticum: Adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. *PROTEOMICS* 2010, *10*, 541-554.

[17] Jennert, K. C., Tardif, C., Young, D. I., Young, M., Gene transfer to Clostridium cellulolyticum ATCC 35319. *Microbiology* 2000, *146 Pt 12*, 3071-3080.

[18] Bayer, E. A., Shimon, L. J. W., Shoham, Y., Lamed, R., Cellulosomes—Structure and Ultrastructure. *Journal of Structural Biology* 1998, *124*, 221-234.

[19] Saxena, S., Fierobe, H.-P., Gaudin, C., Guerlesquin, F., Belaich, J.-P., Biochemical properties of a beta-xylosidase from Clostridium cellulolyticum. *Applied and Environmental Microbiology* 1995, *61*, 3509-3512.

[20] Desvaux, M., Clostridium cellulolyticum: model organism of mesophilic cellulolytic clostridia. *FEMS Microbiol Rev* 2005, *29*, 741-764.

[21] Belaich, A., Parsiegla, G., Gal, L., Villard, C., *et al.*, Cel9M, a New Family 9 Cellulase of the Clostridium cellulolyticum Cellulosome. *Journal of Bacteriology* 2002, *184*, 1378-1384.

[22] Gowen, C. M., Fong, S. S., Exploring Biodiversity for Cellulosic Biofuel Production. *Chemistry & Biodiversity* 2010, *7*, 1086-1097.

[23] Gal, L., Pages, S., Gaudin, C., Belaich, A., *et al.*, Characterization of the cellulolytic complex (cellulosome) produced by Clostridium cellulolyticum. *Applied and Environmental Microbiology* 1997, *63*, 903-909.

[24] Fendri, I., Tardif, C., Fierobe, H.-P., Lignon, S., *et al.*, The cellulosomes from Clostridium cellulolyticum. *FEBS Journal* 2009, *276*, 3076-3086.

[25] Mechaly, A., Fierobe, H.-P., Belaich, A., Belaich, J.-P., *et al.*, Cohesin-Dockerin Interaction in Cellulosome Assembly: A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *Journal of Biological Chemistry* 2001, *276*, 9883-9888.

[26] Vazana, Y., Moraïs, S., Barak, Y., Lamed, R., Bayer, E. A., Chapter twenty-three - Designer Cellulosomes for Enhanced Hydrolysis of Cellulosic Substrates, in: Harry, J. G. (Ed.), *Methods in enzymology*, Academic Press 2012, pp. 429-452.

[27] Tsai, S.-L., DaSilva, N. A., Chen, W., Functional Display of Complex Cellulosomes on the Yeast Surface via Adaptive Assembly. *ACS Synthetic Biology* 2012, *2*, 14-21.

[28] Gelhaye, E., Petitdemange, H., Gay, R., Adhesion and growth rate of Clostridium cellulolyticum ATCC 35319 on crystalline cellulose. *Journal of Bacteriology* 1993, *175*, 3452-3458.

[29] Giallo, J., Gaudin, C., Belaich, J. P., Petitdemange, E., Caillet-Mangin, F., Metabolism of glucose and cellobiose by cellulolytic mesophilic Clostridium sp. strain H10. *Applied and Environmental Microbiology* 1983, *45*, 843-849.

[30] Desvaux, M., Guedon, E., Petitdemange, H., Cellulose Catabolism by Clostridium cellulolyticum Growing in Batch Culture on Defined Medium. *Applied and Environmental Microbiology* 2000, *66*, 2461-2470.

[31] Desvaux, M., Guedon, E., Petitdemange, H., Metabolic flux in cellulose batch and cellulose-fed continuous cultures of Clostridium cellulolyticum in response to acidic environment. *Microbiology* 2001, *147*, 1461-1471.

[32] Dien, B. S., Cotta, M. A., Jeffries, T. W., Bacteria engineered for fuel ethanol production: current status. *Appl Microbiol Biotechnol* 2003, *63*, 258-266.

[33] Guedon, E., Desvaux, M., Petitdemange, H., Kinetic Analysis of Clostridium cellulolyticum Carbohydrate Metabolism: Importance of Glucose 1-Phosphate and Glucose 6-Phosphate Branch Points for Distribution of Carbon Fluxes Inside and Outside Cells as Revealed by Steady-State Continuous Culture. *J. Bacteriol.* 2000, *182*, 2010-2017.

[34] Payot, S., Guedon, E., Cailliez, C., Gelhaye, E., Petitdemange, H., Metabolism of cellobiose by Clostridium cellulolyticum growing in continuous culture: evidence for decreased NADH reoxidation as a factor limiting growth. *Microbiology* 1998, *144*, 375-384.

[35] Guedon, E., Desvaux, M., Petitdemange, H., Improvement of cellulolytic properties of Clostridium cellulolyticum by metabolic engineering. *Appl Environ Microbiol* 2002, *68*, 53-58.

[36] Li, Y., Tschaplinski, T. J., Engle, N. L., Hamilton, C. Y., *et al.*, Combined inactivation of the Clostridium cellulolyticum lactate and malate dehydrogenase genes substantially increases

ethanol yield from cellulose and switchgrass fermentations. *Biotechnology for biofuels* 2012, *5*, 2.

[37] Mingardon, F., Perret, S., Bélaïch, A., Tardif, C., *et al.*, Heterologous Production, Assembly, and Secretion of a Minicellulosome by Clostridium acetobutylicum ATCC 824. *Applied and Environmental Microbiology* 2005, *71*, 1215-1222.

[38] Fierobe, H.-P., Mechaly, A., Tardif, C., Belaich, A., *et al.*, Design and Production of Active Cellulosome Chimeras: Selective incorporation of dockerin-contaning enzymes into defined functional complexes. *Journal of Biological Chemistry* 2001, *276*, 21257-21261.

[39] Lilly, M., Fierobe, H. P., Van Zyl, W. H., Volschenk, H., Heterologous expression of a Clostridium minicellulosome in Saccharomyces cerevisiae. *FEMS yeast research* 2009, *9*, 1236-1249.

[40] Qureshi, N., Blaschek, H. P., Recent advances in ABE fermentation: hyper-butanol producing *Clostridium beijerinckii* BA101. *Journal of Industrial Microbiology & Biotechnology* 2001, *27*, 287-291.

[41] Lütke-Eversloh, T., Bahl, H., Metabolic engineering of Clostridium acetobutylicum: recent advances to improve butanol production. *Current Opinion in Biotechnology*, *In Press, Corrected Proof*.

[42] Bowring, S. N., Morris, J. G., Mutagenesis of Clostridium acetobutylicum. *Journal of Applied Microbiology* 1985, *58*, 577-584.

[43] Liyanage, H., Young, M., Kashket, E. R., Butanol tolerance of *Clostridium beijerinckii* NCIMB 8052 associated with down-regulation of gldA by antisense RNA. *J Mol Microbiol Biotechnol* 2000, *2*, 87-93.

[44] Nicolaou, S. A., Gaida, S. M., Papoutsakis, E. T., Coexisting/Coexpressing Genomic Libraries (CoGeL) identify interactions among distantly located genetic loci for developing complex microbial phenotypes. *Nucleic Acids Research* 2011, *39*, e152-e152.

[45] Gill, R. T., Wildt, S., Yang, Y. T., Ziesman, S., Stephanopoulos, G., Genome-wide screening for trait conferring genes using DNA microarrays. *Proceedings of the National Academy of Sciences* 2002, *99*, 7033-7038.

[46] Blouzard, J.-C., Valette, O., Tardif, C., de Philip, P., Random mutagenesis of Clostridium cellulolyticum using a Tn1545 derivative. *Appl. Environ. Microbiol.* 2010, AEM.02417-02409.

[47] Woodruff, L. B. A., Pandhal, J., Ow, S. Y., Karimpour-Fard, A., *et al.*, Genome-scale identification and characterization of ethanol tolerance genes in Escherichia coli. *Metabolic Engineering* 2013, *15*, 124-133.

[48] Warnecke, T., Lynch, M., Karimpour-Fard, A., Lipscomb, M., *et al.*, Rapid dissection of a complex phenotype through genomic-scale mapping of fitness altering genes. *Metabolic Engineering* 2010, *12*, 241-250.

[49] Bonomo, J., Lynch, M. D., Warnecke, T., Price, J. V., Gill, R. T., Genome-scale analysis of anti-metabolite directed strain engineering. *Metab. Eng.* 2008, *10*, 109-120.

[50] Gall, S., Lynch, M. D., Sandoval, N. R., Gill, R. T., Parallel mapping of genotypes to phenotypes contributing to overall biological fitness. *Metabolic Engineering* 2008, *10*, 382-393.

[51] Borden, J. R., Papoutsakis, E. T., Dynamics of genomic-library enrichment and identification of solvent tolerance genes for Clostridium acetobutylicum. *Appl Environ Microbiol* 2007, *73*, 3061-3068.

[52] Borden, J. R., Jones, S. W., Indurthi, D., Chen, Y., Papoutsakis, E. T., A genomic-library based discovery of a novel, possibly synthetic, acid-tolerance mechanism in Clostridium

acetobutylicum involving non-coding RNAs and ribosomal RNA processing. *Metab Eng* 2010, *12*, 268-281.

[53] Nicolaou, S. A., Gaida, S. M., Papoutsakis, E. T., A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: From biofuels and chemicals, to biocatalysis and bioremediation. *Metabolic Engineering* 2010, *12*, 307-331.

[54] Arneson, N., Hughes, S., Houlston, R., Done, S., Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). *Cold Spring Harbor Protocols* 2008, *2008*, pdb.prot4919.

[55] Johnson, E. A., Madia, A., Demain, A. L., Chemically defined minimal medium for growth of the anaerobic cellulolytic thermophile Clostridium thermocellum. *Applied and environmental microbiology* 1981, *41*, 1060.

[56] Girbal, L., Mortier-Barrière, I., Raynaud, F., Rouanet, C., *et al.*, Development of a sensitive gene expression reporter system and an inducible promoter-repressor system for Clostridium acetobutylicum. *Applied and environmental microbiology* 2003, *69*, 4985-4988.

[57] Harris, L. M., Welker, N. E., Papoutsakis, E. T., Northern, Morphological, and Fermentation Analysis of spo0A Inactivation and Overexpression in Clostridium acetobutylicum ATCC 824. *J. Bacteriol.* 2002, *184*, 3586-3597.

[58] Telenius, H. k., Carter, N. P., Bebb, C. E., Nordenskjo¨ld, M., *et al.*, Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* 1992, *13*, 718-725.

[59] Ng, G., Roberts, I., Coleman, N., Evaluation of 3 methods of whole-genome amplification for subsequent metaphase comparative genomic hybridization. *Diagnostic Molecular Pathology* 2005, *14*, 203-212.

[60] Bimboim, H. C., Doly, J., A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* 1979, *7*, 1513-1523.

[61] Cui, G.-z., Hong, W., Zhang, J., Li, W.-l., *et al.*, Targeted gene engineering in *Clostridium cellulolyticum* H10 without methylation. *Journal of microbiological methods* 2012, *89*, 201-208.

[62] Gehin, A., Gelhaye, E., Raval, G., Petitdemange, H., Clostridium cellulolyticum Viability and Sporulation under Cellobiose Starvation Conditions. *Applied and environmental microbiology* 1995, *61*, 868-871.

[63] McNeil, B., Harvey, L., *Practical Fermentation Technology*, Wiley 2008.

[64] García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., *et al.*, Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012, *28*, 2678-2679.

[65] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., *et al.*, The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bmc Bioinformatics* 2008, *9*, 386.

[66] Jennert, K. C., Tardif, C., Young, D. I., Young, M., Gene transfer to Clostridium cellulolyticum ATCC 35319. *Microbiology* 2000, *146*, 3071-3080.

[67] Marvin, D. A., Filamentous phage structure, infection and assembly. *Current opinion in structural biology* 1998, *8*, 150-158.

[68] Ren, Z., Ward, T. E., Logan, B. E., Regan, J. M., Characterization of the cellulolytic and hydrogen-producing activities of six mesophilic Clostridium species. *Journal of Applied Microbiology* 2007, *103*, 2258-2266.

[69] Li, Y., Xu, T., Tschaplinski, T. J., Engle, N. L., *et al.*, Improvement of cellulose catabolism in Clostridium cellulolyticum by sporulation abolishment and carbon alleviation. *Biotechnology for biofuels* 2014, *7*, 25.

[70] Tsay, J. T., Rock, C. O., Jackowski, S., Overproduction of beta-ketoacyl-acyl carrier protein synthase I imparts thiolactomycin resistance to Escherichia coli K-12. *Journal of Bacteriology* 1992, *174*, 508-513.

[71] Skovran, E., Crowther, G. J., Guo, X., Yang, S., Lidstrom, M. E., A systems biology approach uncovers cellular strategies used by Methylobacterium extorquens AM1 during the switch from multi-to single-carbon growth. *PloS one* 2010, *5*, e14091.

[72] Lehnherr, H., Maguin, E., Jafri, S., Yarmolinsky, M. B., Plasmid addiction genes of bacteriophage P1: doc, which causes cell death on curing of prophage, and phd, which prevents host death when prophage is retained. *Journal of molecular biology* 1993, *233*, 414-428.

[73] Cui, G.-Z., Zhang, J., Hong, W., Xu, C., *et al.*, Improvement of ClosTron for successive gene disruption in Clostridium cellulolyticum using a pyrF-based screening system. *Appl Microbiol Biotechnol* 2014, *98*, 313-323.

[74] Tyurin, M. V., Desai, S. G., Lynd, L. R., Electrotransformation of Clostridium thermocellum. *Applied and Environmental Microbiology* 2004, *70*, 883-890.

[75] Freedman, B. G., Zu, T. N. K., Wallace, R., Senger, R. S., Raman spectroscopy characterizes 1-butanol tolerant *E. coli* phenotypes resulting from genomic DNA library enrichment. *Submitted* 2014.

[76] Freedman, B. G., Senger, R. S., Thermodynamic design of degenerate oligonucleotide primers for whole genome DNA amplification with reduced bias for use in DNA libraries *Submitted* 2014.

Chapter 5

The Feedback Inhibition of Transcription for Substrain Selection (FITSelect) genetic circuit to couple cell growth rate to non-essential metabolite production

ABSTRACT

A novel genetic circuit called "Feedback Inhibition of Transcription for Substrain Selection (FITSelect)" was created in this research to artificially tie the growth rate of a cell to its production rate of L-arginine. FITSelect ensures that high producing cells grow at a faster rate than low producers. Thus, combinatorial (i.e., randomized) metabolic engineering strategies can be carried out and growth competition can be used to isolate productive strains. L-arginine was used as a target for this proof-of-concept, but the FITSelect circuit can be altered easily to enrich for production of hundreds of different high-value chemicals. The FITSelect circuit consists of an L-arginine responsive promoter that controls a *luxR* gene. High L-arginine concentrations repress *luxR* expression. LuxR and exogenously added N-(β-Ketocaproyl)-L-homoserine lactone (OHHL) activate expression of a cell death gene (*ccdB*) fused with a *lacZ* reporter. Thus, cells without high concentrations of L-arginine ultimately activate the *lacZ-ccdB* fusion gene, which induces apoptosis. Here the construction and testing of the FITSelect circuit with exogenously supplied L-arginine and other amino acids is described. Very good response by all parts of the circuit and selectivity for L-arginine was observed. Expression of a genomic DNA library in a culture containing the FITSelect circuit was performed to identify DNA fragments that could improve L-arginine production. However, enrichment experiments yielded similar fragments for

induced and un-induced cultures. This leads us to believe that, though the circuit functions in controlled conditions, additional fine-tuning of individual circuit components is still needed to selectively enrich for strains with high L-arginine production.

INTRODUCTION

The FITSelect gene circuit

A synthetic gene circuit was derived in this research to aid a metabolic engineering process for the production of L-arginine from *E. coli*. The circuit functions to provide a growth and/or fitness advantage to individual cells engineered to have higher internal L-arginine concentrations. Thus, the circuit serves as a screening mechanism that allows growth competition to select for the highest L-arginine producer. The circuit is based on the two-component regulatory system using the quorum sensing LuxR protein from *Vibrio fischeri* and the control of cell death protein, CcdB [1]. The system is called the "Feedback Inhibition of Transcription for Substrain Selection (FITSelect)" and makes use of L-arginine sensitive promoters, regulated by the repressive feedback of L-arginine, that control production of the *ccdB* cell death gene. The full FITSelect circuit is shown in Figure 5-1.

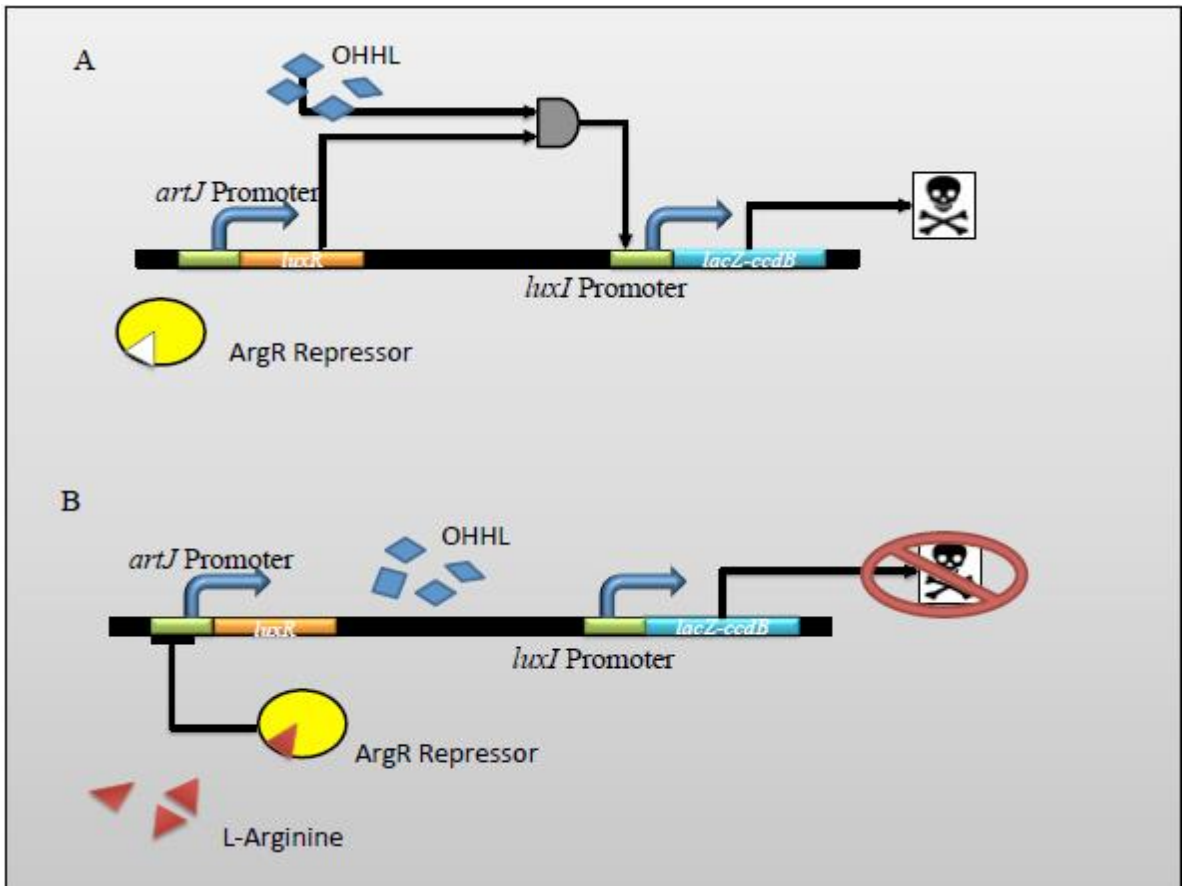


Figure 5-1. FITSelect circuit model. (A) Function of circuit in absence of L-arginine. The ArgR genomic repressor does not bind to the *artJ* promoter and transcription of *luxR* occurs. Exogenously supplemented N-(3-oxohexanoyl)-l-homoserine lactone (OHHL) will combine with the LuxR to activate transcription of the *lacZ-ccdB* gene. The *ccdB* portion of the gene retains the ability to inhibit DNA gyrase and inhibits growth and replication. In growth conditions with excess L-arginine (B), the ArgR repressor is bound to L-arginine, allowing it to bind to the promoter and inhibit *luxR* mRNA operator and decreases the production of LuxR and ultimately the *lacZ-ccdB* gene. Other plasmids used in this study have modified promoters in place of the *artJ* promoter or contain *mCherry* in place of *luxR*.

The initial goal was to characterize the performance of L-arginine sensitive promoters using an mCherry reporter fluorescent protein [2] and previously characterized promoters in the L-arginine synthesis pathway [3, 4]. Prior studies have taken advantage of feedback regulation to create an L-lysine sensor by fusing the gene coding for the green fluorescent protein (GFP) to the lysine-responsive *dapD* gene promoter [5]. We tested the *artJ* and *argF* promoters for their responses to L-arginine concentrations and then examined their function when transcribing *luxR,* the quorum sensor activator (along with N-(β-Ketocaproyl)-L-homoserine lactone (OHHL)) of the *lux* operator controlling *ccdB* gene expression. The FITSelect circuit was built on the hypothesis that cells harboring higher L-arginine levels will have activated their L-arginine biosynthesis repressor proteins. When activated by L-arginine, the repressors block transcription from several promoters (e.g., *argF* and *artJ*) in the L-arginine biosynthesis pathway. In FITSelect, both the *argF* and *artJ* promoters were put in control of the *luxR* gene, with the *artJ* promoter eventually showing a favorable response (to be discussed). LuxR and OHHL together activate transcription from the *luxI* promoter. This promoter controls expression of the *ccdB* cell death gene. As shown in Figure 5-1, the *lux* system is used to control activation of the circuit and delay the response of cell death, making the circuit activity controllable through the amount of OHHL analogues added. Thus, the goal of FITSelect is to eliminate cells that produce low levels of L-arginine. If FITSelect is expressed in cells engineered by chemical or UV mutagens, transposon-based plasmids, or genomic DNA libraries, the circuit is designed to enable enrichment (i.e., survival and growth) of only cells with mutations or added genes that increase internal L-arginine concentrations. Thus, FITSelect is designed as a gene circuit aimed at solving a critical problem in metabolic engineering: tying the growth rate of a cell to its production of a target chemical. With this, the "survival-of-the-fittest" challenges will produce a culture capable

of high target product formation rates. In this proof-of-concept design, L-arginine was chosen as the target product due to the availability of characterized L-arginine responsive promoters. Variations of FITSelect are possible for several other chemical targets.

L-arginine regulatory mechanisms

L-arginine is predominantly produced from fermentation of glucose with high oxygen and ammonia requirements. It is formed directly from L-ornithine and citrulline, which derive from the L-glutamic acid pool [6]. L-arginine production is regulated by the ArgR regulatory protein. This repressor protein is 17 kD and exists at 40-200 protein copies per cell, functioning predominantly as a hexamer and occasionally a trimer. ArgR has binding sites for both the ligand L-arginine and 18bp palindromic *arg* box motifs that precede genes encoding arginine synthesis enzymes (*argACDEFGI)*, arginine transport (*artIJMQ)*, histidine transport (*hisJMQ)*, carbamoylphosphate synthesis (*carAB*) and even inhibits its own production [7]. As the *arg* box makes up the promoter and transcriptional start site of the genes it precedes, L-arginine dependent regulation is thought to function as a feedback response, sterically blocking attachment of the polymerase to prevent the cell from generating L-arginine biosynthesis enzymes when excess exists. Binding of ArgR to the *arg* box represses genes downstream 10-350x more when ArgR is also bound to L-arginine with the level of control varying depending on the sequence, spacing, and number of *arg* boxes present. This mechanism of regulation is thought to conserve cellular energy and protein production without utilizing attenuation [8].

L-arginine as a value-added product

L-arginine currently sells for ~$200/kg, and its use as a nutritional supplement is currently popular. It is believed to increase muscle mass [9] and potentially reduce adipose

(fatty) tissue (as demonstrated in rats, pigs) and type II diabetes in humans by regulating lipogenic and lipolytic genes [10]. Another important global market for L-arginine production is as a supplement for animal feed. In this capacity, it has been shown to increase average daily weight gain [11] and feed intake in piglets [12]; milk yield in cows [13]; increased weight gain, feed intake, and gain/feed ratio in broiler chickens [14]; and improved egg production, egg weight, and stock survivability in egg layers [15].

Prior efforts to increase L-arginine production

Industrial production of total amino acids exceeds 2 million metric tons per year and most amino acids are now produced via fermentation with *Corynebacterium glutamicum* [16, 17], though proteolysis is still utilized in some countries [6] and chemical synthesis from L-ornithine remains possible if not cost effective. Increased production of L-arginine in fermentation has been achieved through (i) mutagenesis and antimetabolite selection, (ii) targeted mutagenesis, and (iii) metabolic pathway re-design [17]. As L-glutamate is a common starting point in the L-arginine synthesis pathway, the strong glutamate producer *Corynebacterium* (*Brevibacterium*) is often used in strain enrichment studies. The use of mutagen N-methyl-N'-nitro-N-nitrosoguanidine (NG) or diethyl sulfate (DES) can cause spontaneous mutations in cells. Treatment with NG creates mutants that are screened by their activity towards anti-metabolites, or analogues to amino acids that compete for activity of their native counterpart. Common anti-metabolites or antagonists include arginine hydroxamate, cadaverine, homoarginine, and 2-thiazolealanine, which are believed to be false feedback inhibitors, false co-repressors, or arginyl-transfer RNA synthetase inhibitors [18]. In one such study, *Brevibacetrium flavam* (ATCC 14067) was subjected to multiple rounds of irradiation, NG, and DES treatment followed

by selection for resistance to increasing titers of 2-thiazolealanine. Mutants were manually screened and the organisms with the highest L-arginine production levels underwent further enrichment to generate a strain capable of accumulating 25.3g/L, which was further increased to 34.8g/L in a 13% glucose broth after medium optimization [19]. Another targeted engineering approach used the wild-type *C. glutamicum* in which two bases in the *argB* gene were altered to eliminate end product feedback inhibition. This was followed by elimination of the *argR* repressor gene to remove the global L-arginine regulator. The resulting strain produced L-arginine at 52.8 g/L, slightly less than the classical strain A-27 at 65.7 g/L [20]. From an industrial perspective, culturing and aeration considerations can complicate fermentation. Improvements to this process, such as adding in a hemoglobin from *Vitreoscilla* to a *Corynebacterium creanatum* to improve oxygen uptake resulted in a 10.5% increase in biomass and 17.3% increase in L-arginine in a batch fermentation [21].

The role of FITSelect

Here, we have developed a new genetic circuit-based approach to further select for cells that are high-producers of L-arginine. While it is true that significant efforts have been made previously to increase L-arginine production, additional gains are possible. In addition, we have estimated the mechanisms used in the FITSelect system make it applicable to over 100 additional chemicals with commercial value. In order to be compatible with FITSelect, a target chemical must activate a repressor protein that can inhibit transcription by a promoter. This promoter is then oriented in front of the *luxR* gene of FITSelect, and this adapts the circuit for the enrichment based on another chemical of interest. In this research, the FITSelect circuit was assembled,

tested, and used to select for high L-arginine producers from *E. coli* cells transformed with a DNA library.

MATERIALS AND METHODS

Cell lines, chemicals, and media

N-(β-Ketocaproyl)-L-homoserine lactone (OHHL) was obtained from Sigma-Aldrich (St. Louis, MO). Other chemicals were obtained from ThermoFisher Scientific (Waltham, MA). Enzymes were obtained from New England Biolabs (Ipswitch, MA) unless otherwise noted. The cell lines used along with plasmids used and constructed are listed in Table 5-1. *E. coli* NEB 5-α was used for circuit expression, *E. coli* MG1655 for the source of genomic library fragments, and One Shot® *ccdB* Survival™ 2 $T1^R$ for propagation of the pDEST14 destination vector (Life Technologies; Grand Island, NY). The *ccdB* Survival cell line is resistant to the toxicity of CcdB and was also used to examine *ccdB* expression, in addition to FITSelect circuit propagation. Both LB and M9 minimal media were supplemented with 2 mg/L thiamine and used for growth and propagation of the clones at 37°C and 200 rpm shaking. L-arginine, L-ornithine, and L-histidine were added as 100 mM solutions. OHHL was added from a 10 mM stock stored at -40°C.

FITSelect circuit constructions

Gblocks and primers (Integrated DNA Technologies; Coralville, IA) used are listed in Supplementary Appendix Table 7-4. The *luxR* gene was provided in plasmid pJE202 [22]. The *luxI* promoter and the *lacZ-ccdB* gblock were amplified with primer sets LuxR_F/LuxR_R and Fslux_gblockF/Fslux_gblockR (Table 7-4) using Q5 High Fidelity Polymerase (New England Biolabs), purified, combined at equimolar ratios with the FS_osip_F/Osip_AatII primers, and

amplified with Q5 polymerase to produce the working FITSelect circuit via Splicing by Overlap Extension PCR [23]. Though both unmodified *ccdB* was used previously, addition of the catalytic *lacZ* subunit modulates the potency of the toxin and enables usage of the β-galactosidase activity to monitor production of the toxin [1]. The circuit was digested with AatII and BamHI into ligated with T4 DNA Ligase (New England Biolabs) into similarly digested and dephosphorylated pACYC177 (New England Biolabs). The plasmid was linearized with phosphorylated primers FitSel_EcoRI/Fitsel_XbaI and rrnBT1T2_F/rrnBT1T2_EcoRI (Table 7-4) to introduce the EcoRI and XbaI restriction sites upon circularization. The plasmid was further amplified with primers LuxR_NdeI/FS_bbone_R to introduce an NdeI cut site, followed by PCR with dephosphorylated primers KanR_F/Gblock_AatII_R to remove the ampicillin cassette. The terminator sequence *rnbT1T2* was amplified from pBad24 [24], using primers Stop_F/Stop_R (Table 7-4) and digested with EcoRI and XbaI, and cloned into a similarly digested vector, producing FITSelectRRNB. FiteselectCherry was produced by digesting FitselectRRNB with NdeI and BamHI and ligating to the similarly digested PCR product generated from amplifying plasmid pKDL071 with primers Cherry_ndeI/Cherry_bamHI (Table 7-4). Promoter *ArtJweak* was amplified using primers ArtJ_weakRBS/ArtJ_Eco_F from *E. coli* MG1665 genomic DNA targeting the promoter to gene *artJ*. The 160 bp fragment was inserted into FitSelectRRNB and FITSelectCherry using NdeI and EcoRI to generate FSArtJWeakLux and FSArtJWeakcherry. Promoter *ArtJweakSD* containing a mutation in the -35 region of the *artJ* promoter[3] was ordered as a gblock and digested with NdeI and EcoRI into similarly digested FitSelectRRNB and FITSelectCherry to generate FSArtJSDLux and FSArtJSDCherry.

Table 5-1. List of strains and plasmids used in this study.

Strains	**Genotype**	**Source**
E. coli NEB5-α	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	New England Biolabs
E. coli K12 MG1665	Wild-type	Personal Collection
E. coli ccdB Survival™ 2 T1^{R}	*F-mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 araΔ139 Δ(ara-leu)7697galU galK rpsL (StrR) endA1 nupG fhuA::IS2*	Life Technologies
Plasmids	**Components**	**Source**
pJE202	Lux cassette	Stevens et al. 1994
pBAD24	*araBAD* promoter, rrnB T1T2 transcription terminator	Guzman et al. 1995
pKDL071	mcherry, gfpmut3b	Personal Collection
pACYC177	Amp^{R}, Km^{R}, p15A ori	Life Technologies
pDest14	T7 promoter, Gateway cloning cassette, pBR322ori, Amp^{R}	Life Technologies
FITSelect	*lacz-ccdB, luxR,* Amp^{R}, Km^{R}, p15A ori	This study
FITSelectRRNB	*lacz-ccdB, luxR, rrnb*T1T2, Km^{R}, *argF* prometer, p15A ori	This study
FITSelectCherry	*lacz-ccdB, rrnb*T1T2, Km^{R}, *argF* promoter, mCherry, p15A ori	This study
FSArtJWeak_Cherry	*lacz-ccdB, rrnb*T1T2, Km^{R}, *artJ* promoter, mCherry, p15A ori	This study
FSArtJWeak_Lux	*lacz-ccdB, luxR, rrnb*T1T2, Km^{R}, *artJ* promoter, p15A ori	This study
FSArtJSD_Cherry	*lacz-ccdB, rrnb*T1T2, Km^{R}, *artJ** promoter, mCherry p15A ori	This study
FSArtJSD_Lux	*lacz-ccdB, luxR, rrnb*T1T2, Km^{R}, *artJ** promoter, p15A ori	This study

Cell culture measurements

FitSelectCherry, FSArtJSDCherry, FSArtJWeakcherry, FSArtJSDLux and FSArtJWeakLux were grown from stock cultures in minimal media. A total volume of 200 µL of M9 with 50 mg/L kanamycin was added to each well of a covered 96-well round-bottom plate (Corning) and 5 µL of cells were added to an initial OD_{600} of 0.05. L-arginine and OHHL were added at varying concentrations to assess the response of the *artJ* and *argF* promoters to inducers and the basal expression of *lacz-ccdB* with OHHL but no *luxR*. Cells were measured in a BioTek® Synergy H4 well plate reader. OD was measured at 600 nm and fluorescence at 584 nm excitation and 612 nm emission. Covered 96-well plates were sealed with Parafilm "M" to reduce evaporation. Samples were performed in triplicate wells. Data was analyzed using custom R scripts and the Grofit analysis program [25].

Determining LacZ-CcdB activity

The LacZ-CcdB fusion protein under control of the *luxI* promoter was measured as cells were grown in minimal or L-arginine rich M9 media. Overnight cultures of FSArtJWeakLux in minimal M9 media were normalized to an OD_{600} of 0.05 and grown overnight in 5 mL of 0, 50, and 250 µM L-arginine in triplicate with and without 10^{-6} M OHHL. Activity of the β-glactisidase fusion was performed exactly as described [1] using the FluoReporter® LacZ/Galactosidase Reporter Kit (Life Technologies).

DNA library construction

Genomic libraries were produced through degenerate oligonucleotide primed PCR (DOP-PCR) amplification of genomic DNA from *E. coli* K12 MG1655 extracted with a Generation

Capture Column (Valencia, CA). The amplification was performed using primers M1, M2, M4, M5, and Rand3 (Table 7-1) in separate, 50 µL reactions using Long-Amp *Taq* Polymerase (New England Biolabs) as described previously (Chapters 2 and 3). Purified reaction mixtures were used directly with PCR8-TOPO-TA vectors to generate 20,000 mutants and combined with pDEST14 with Clonease II (Invitrogen) to generate another 23,000 mutants in *E. coli* NEB 10-beta competent cells growing on ampicillin alone. Plasmid DNA was isolated by midiprep and was combined with *E. coli* NEB 5-α cells harboring the FSArtJWeakLux plasmid, made competent by the $CaCl_2$ method [26], to generate 3 $x10^6$ colonies grown in LB media with ampicillin and kanamycin selection. The library was also introduced into *E. coli* NEB 5-α cells without the FITSelect circuit to generate 6 $x10^6$ colonies grown in LB with ampicillin as the sole antibiotic.

DNA library screening

Freshly transformed cells containing both the FITSelect containing plasmid on the p15A origin and the library inserts on the pBR322 origin of replication were diluted to an OD_{600} of 1.0. A volume of 100 µL was added into (i) a culture with 10^{-6} M OHHL and (ii) a culture without any OHHL in 100 mL M9 media with 50 mg/L kanamycin and 100 mg/L ampicillin in a 500 mL flask. In order to reduce the effects of FITSelect circuit mutations, which were found to negate the FSArtJWeak_Lux circuit (sometimes after one 24 hour outgrowth), minipreps were performed on both cultures and frozen, competent *E. coli* NEB 5-α harboring the FSArtJWeak_Lux plasmid were transformed with the minipreps from the previous outgrowth. After a 1 hour recovery in SOC medium, cells were pelleted, washed, and added directly to fresh M9 medium with kanamycin and ampicillin as before. After 24 hours, cells were plated on solid

LB with kanamycin and ampicillin and screened with PCR and the universal T7 oligonucleotide primers. Unique mutants were miniprepped, transformed into *E. coli* NEB 5-α cells and selected on ampicillin to generate a cell with only the library expression plasmid without the presence of the FITSelect circuit to test the function of the enriched DNA fragment(s).

L-arginine assay

-Arginine concentrations were measured from cell pellets of overnight cultures grown in M9 minimal media with 100 mg/L ampicillin. The pellet was washed twice with phosphate-buffered saline and subjected to the Sakaguchi reaction with 8-hydroxyquinoline as recently described [27].

RESULTS

The L-arginine responsive promoters

To assess the responsiveness of the synthetic circuit to L-arginine concentrations, the *mCherry* [2] fluorescent reporter was fused to the *E. coli argF*,*artJ*, and modified *artJ* (henceforth *artJSD*) gene promoters, utilizing the native ribosome binding site of the *E. coli* genes but modified with a CAT replacement at the -3 to -1 site to accommodate a NdeI restriction site. The results in Figure 5-2 show how each of the cell lines fluoresce in response to L-arginine. Plasmid FITSelectCherry did not show the expected repression of mCherry when grown in the presence of inhibitory L-arginine, nor did it exhibit changes in growth rate when the *mCherry* gene was replaced with *luxR* (Figure 7-6 of the Supplementary Appnedix). This signaled that the *argF* promoter was not adequately responsive to L-arginine concentrations. Fluorescent intensity from expression of mCherry with the *artJ* promoter and *artJSD* mutated

promoter decreased in intensity as expected [8] (Figure 5-2) in both concentrations of OHHL used. However, the OHHL was not expected to affect the function of the L-arginine responsive promoter without luxR (Figure 5-1), and this confirmed the hypothesis. Further, the promoters were tested using thermodynamically designed ribosome binding site (RBS) strengths, over a 10^4-fold activity range [28]. Initial studies indicated that only the weakest of the RBS designs enabled differential expression of mCherry (data not shown).

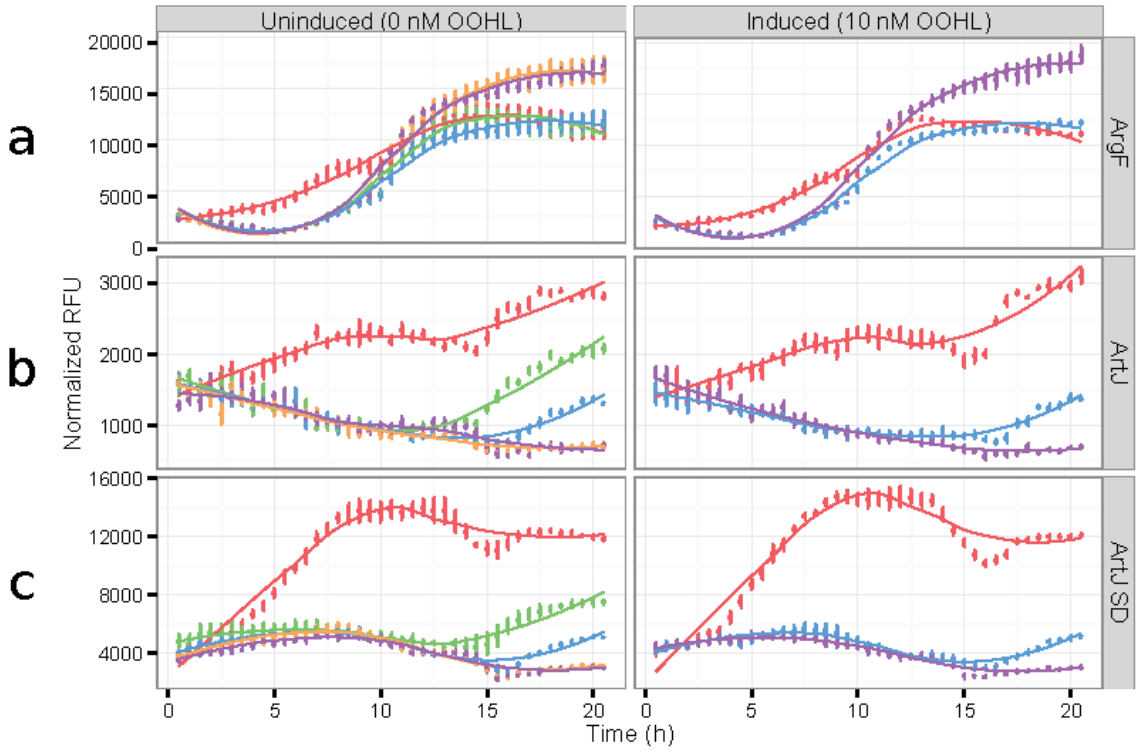


Figure 5-2. Measurement of mCherry fluorescent proteins in response to arginine concentration. Normalized measurements were made for A)FSArgFCherry B)FSArtJCherry, and C)FSArtJSDCherry at A584/612 and OD600 at 0 (red), 50 (yellow) 100 (green) 250 (blue), or 500(purple) µM L-Arginine. OHHL was uninduced (Left) or induced to 10nM (Right) for each sample.

In agreement with results from Caldara et al. [3], the SD mutation increased overall mRNA transcription, resulting in almost a 5-fold increase in activity at 0 µM L-arginine in the circuit with and without the mutation. After 20 h of growth, the FSArtJWeak_Cherry and FSArtJSD_Cherry strains both showed 600% increase in activity between growth in minimal media and in media with L-arginine at 500 µM. Additionally, the induction of the circuit with OHHL without the *luxR* gene showed minimal effect on the cell, even when the *lacZ-ccdB* gene

was expressed in the *ccdB* Survival cell line. After 20 hours, the fluorescence increased for samples with higher concentrations of L-arginine. This is possible if the L-arginine in the media was consumed over time, leading to escape of the argR repressor and reactivation of *mCherry* transcription.

Cell death gene expression

To verify that the *ccdB* cell death gene was expressed, the activity of the LacZ portion of the LacZ-CcdB cell death fusion protein was measured using a β-galactosidase assay (Figure 5-3A). The *lacZ-ccdB* gene was under control of the *luxI* promoter that required the presence of both LuxR and OHHL to be active. It exhibited minimal activity without induction from OHHL, and it was 16x more active in the presence of 10^{-6} M OHHL. The concentration of LuxR was varied by altering the concentration of L-arginine added to the culture. This was found to alter the activity of the *lacI* promoter given a constant OHHL concentration. With L-arginine concentrations of 25 μM, the expression of the cell death gene was reduced 30%. The circuit ceased to show any further change in responsiveness as the L-arginine concentration increased from 25 μM to 250 μM. The cellular concentration of L-arginine is not expected to match the concentration in the bulk media as the cell rapidly produces and consumes free L-arginine to synthesize protein. The internal level of L-arginine is known to increase depending on the exogenous concentration in the media, ranging from 140 μM in minimal media to 1500 μM when the media is supplemented with 570 μM L-arginine [29].

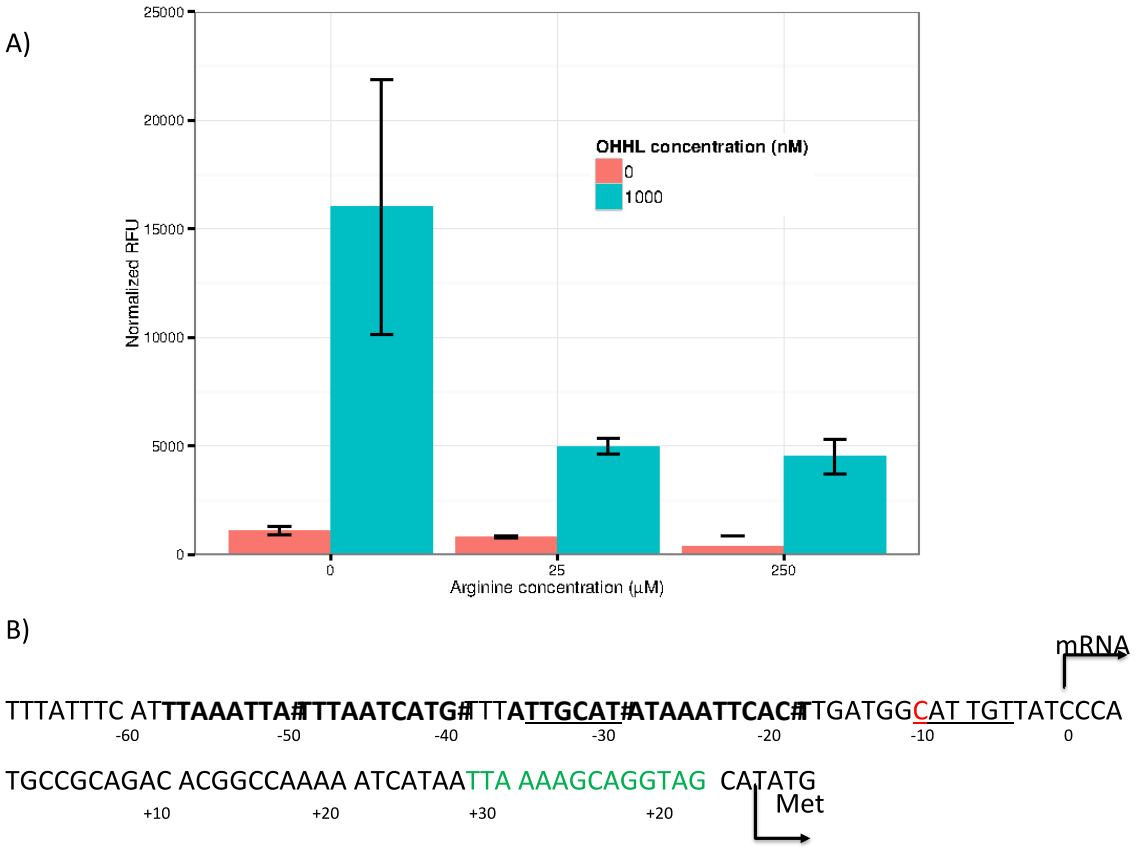


Figure 5-3. (A) β-galactosidase assay of FSArtJWeak. Absorbance at 594 nm was used to determine the activity of the LacZ-CcdB fusion. OHHL induction was performed at 0 M (red) or 10^{-6} M (blue) in M9 minimal media with L-arginine supplementation as indicated. Triplicate samples were performed and error bars represent one standard deviation. (B) Structure of the *artJ* modified promoter in FSArtJWeak plasmids. The mRNA transcription start site and translational start site are shown. Engineered ribosome binding site (green) and Shine-Delgarno consensus mutation in FSArtJSD (red, C->T) are labeled. Bold letters indicate the dual binding domains of the *arg*-box and underlined regions are the -35 and -10 RNA polymerase binding sites.

Assessing FITSelect circuit function

Continuous monitoring of cells containing FSArtJWeak_Lux and FSArtJSD_Lux constructs of FITSelect demonstrated the effect of varying both L-arginine and OHHL concentrations on the growth rate of cultures. This is shown in Figure 5-4. Increasing the concentration of the inducing agent, OHHL, caused a gradual decline in the overall cell density for the FSArtJSD_lux culture but not as much for the ArtJ wild-type promoter. While un-induced cultures showed no response to L-arginine, the induced cultures, starting at 10^{-7} M OHHL, showed the effects of CcdB toxicity when grown under minimal L-arginine levels. The effect of OHHL and L-arginine on growth rate are shown in Figure 5-5. Just as maximal cell density was inhibited, growth rate decreases under exposure to inducing agent, but only in cultures with minimal L-arginine. For the circuit to have maximum utility, the change between the growth rates without additional OHHL should be minimal. Once OHHL is added, the growth rates would ideally decrease inversely to the amount of arginine added. A steeper slope would indicate the greatest dynamic range between over and under-producers of L-arginine

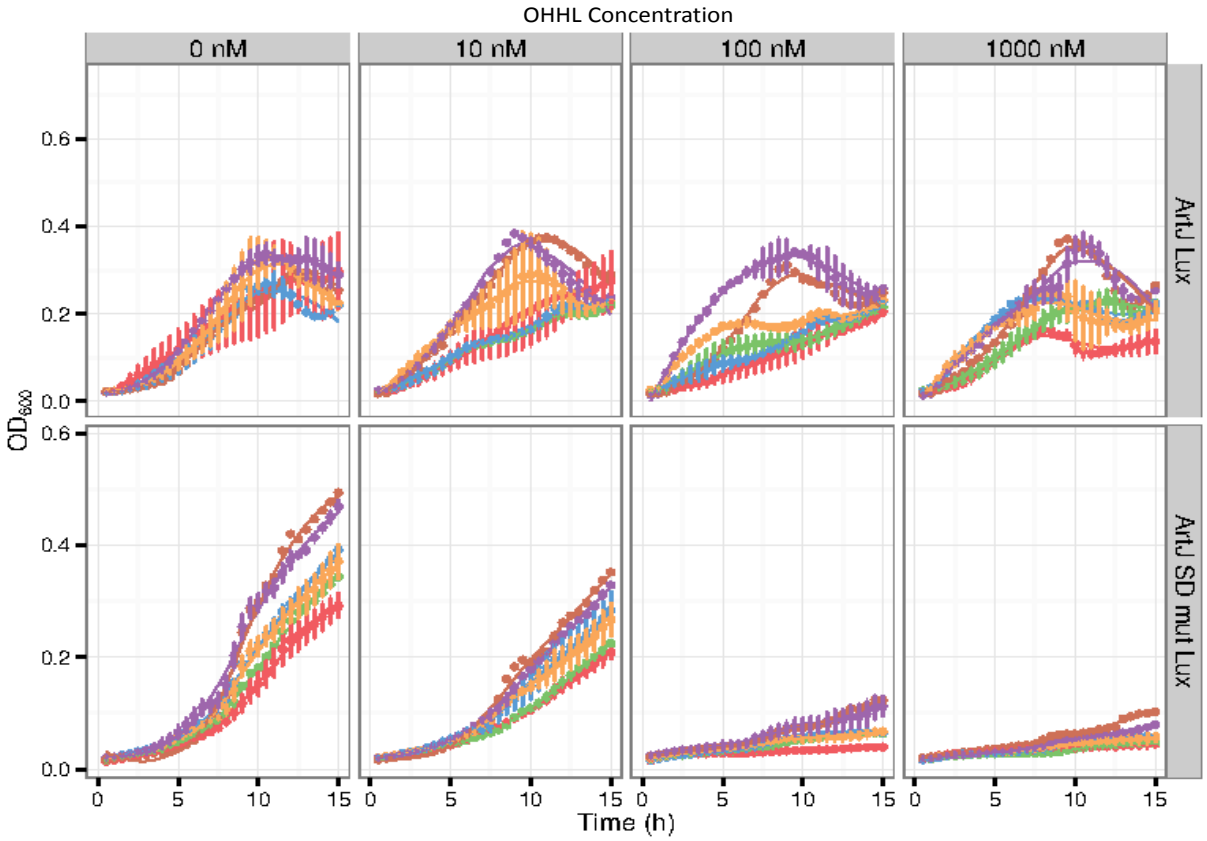


Figure 5-4. Growth of E. coli NEB5α cells harboring a FITSelect genetic circuit grown in M9 minimal media with varying levels of arginine and OHHL. (Top) Growth of FSArtJWeak_Lux and (Bottom) FSArtJSD_Lux at 0 (red) 25 (blue) 50 (green), 100(yellow), 250(purple), and 500 (brown) µM L-arginine added to the media. Concentration of OHHL increases from left to right at 0, 10, 100, and 1000nm. Each points is an average of duplicates and all samples were grown simultaneously in a 96-well plate.

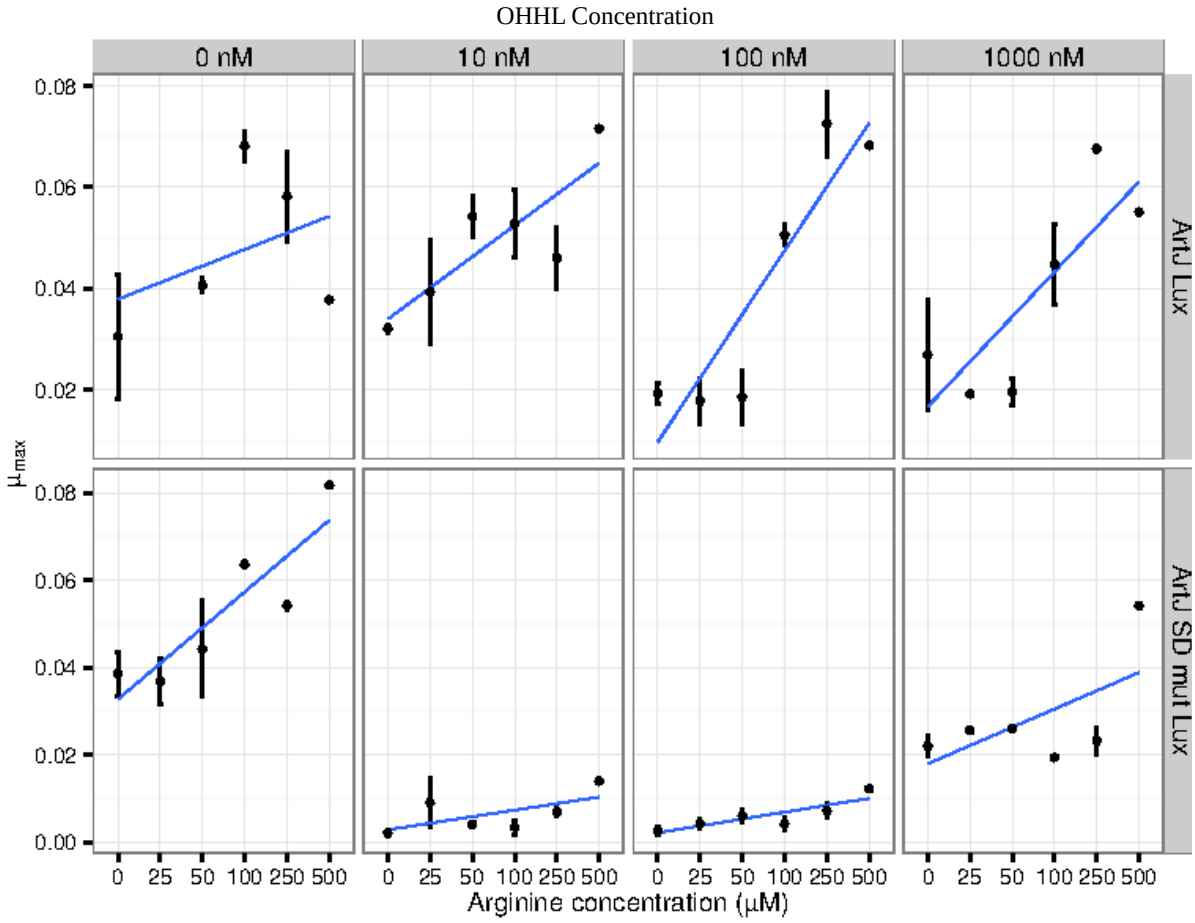


Figure 5-5. Maximum growth rate of *E. coli* NEB5α cells harboring a FITSelect genetic circuit. Cells were grown in M9 minimal media with varying levels of arginine and OHHL. Growth of FSArtJWeak_Lux (TOP) and FSArtJSD_Lux (BOTTOM). Concentration of OHHL increases from left to right at 0, 10^{-8}, 10^{-7}, and10^{-6} M. Each points is an average of four replicates and one error bar represents standard deviations.

To verify that the circuit was responding to L-arginine and not just feeding off the extra carbon and nitrogen of the supplemented media, we examined growth while substituting L-ornithine and L-histidine for L-arginine. L-Ornithine is a necessary precursor to form L-arginine from L-glutamate, and its supplementation was expected to increase free L-arginine pools. L-Histidine is a positively charged, nitrogenous amino acid similar to arginine but not involved in

the L-arginine biosynthesis pathway [3]. Results of the FITSelect circuit function on culture growth when supplementing the media with L-ornithine and L-histidine in place of L-arginine are shown in Figure 5-6. When no OHHL was present (no *lacZ-ccdB* expression), the cultures performed similarly regardless of the amount of endogenous arginine. When the FITSelect circuit was activated with a constant amount of OHHL, a significant decrease in biomass accumulation and culture growth rate was observed (Figure 5-6). This decrease was most pronounced for the cultures supplemented with L-histidine, where even feeding differences did not account for the similar behavior of the 0 and 200 µM cultures. The L-arginine response was largely protected from the effects of OHHL, achieving similar cell density and growth rate as the un-induced sample when protected with high concentrations of 100 or 200 µM L-arginine. The L-ornithine fed cultures also showed a partial recovery in OD_{600} with at least 25 µM supplementation.

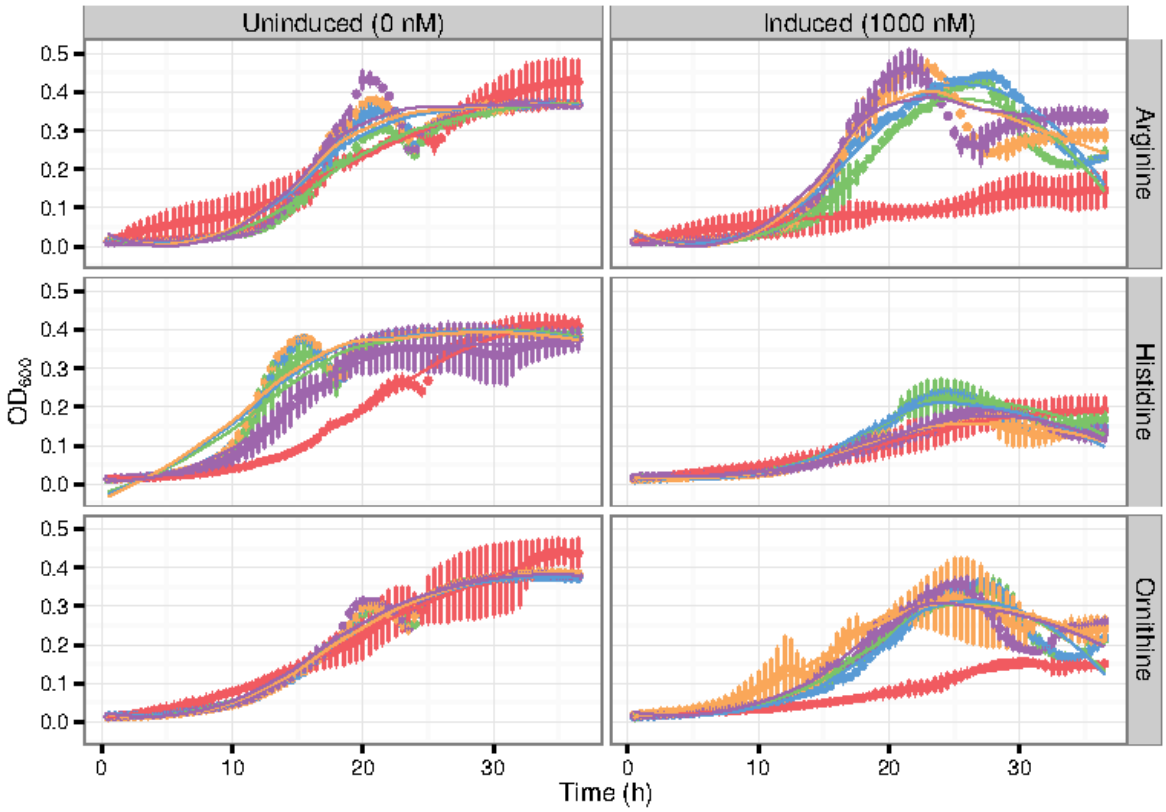


Figure 5-6. Growth of cultures in minimal media harboring the FITSelect circuit on the FSArtJWeak_Lux plasmid with response to L-arginine (top), L-histidine (middle), and L-ornithine (bottom). The supplemented concentrations used were: 0 (red), 25 (green), 50 (blue), 100(orange), and 200µM (purple) of the appropriate amino acid (Left). OHHL was added to 10^{-6}M to activate the FITSelect circuit (Right). Samples are the average of three replicates with standard error bars

Metabolic engineering with DNA library enrichment

Once the FITSelect circuit was shown to respond to L-arginine concentrations by impairing cell growth rate, a genomic DNA library constructed from the *E. coli* MG1665 genome was generated using the DOP-PCR method [30] to produce approximately 24,000 unique DNA fragments of average size of 900 bp (n=20). This DNA library was co-expressed in

cells containing the FSArtJWeak_Lux plasmid and the dual-plasmid cells were selected using the antibiotics kanamycin and ampicillin. When grown in minimal M9 medium with no exogenous L-arginine source and induced with OHHL, the culture selected for DNA library fragment-bearing plasmids that improved overall cell fitness. A genomic DNA library fragment of 2.5 kb was predominantly enriched in 9 of 10 colonies tested in OHHL induced and un-induced enrichments. The DNA fragment contained full coding regions of *purB* and *hfld*, with the terminal end of the *mnmA* tRNA 2-thiouridylase. Two other unique inserts were found to each of the induced and un-induced libraries: the *galR* and *lysA* fragments and *rhsE* pseudogene fragment respectively (Table 5-2). An analysis of L-arginine content of the cultures (Figure 5-7) showed little difference between substrains selected in each enrichment, indicating enrichment may not have occurred with L-arginine production as the sole selective pressure.

Table 5-2. Classification of enriched library fragments co-expressed with FITSelect and isolated from the OHHL induced culture (Positive Selection), the uninduced culture (Negative Selection) or found in both (Neutral Selection).

	Positive Selection	**Neutral Selection**	**Negative Selection**
Library Location	U00096.3: 2977356-2978276	U00096.3: 1190528-1192985	U00096.3: 1527368-1528635
Included genes	GalR, LysA	mnmA, hflD, purB	RhsE
Gene Description	Galactose-inducible repressor of galETK operon, diamionpimelate decarboxylase	tRNA(Gln, Lys, Glu) U34 2-thiouridylase, lysogenization regulator, Adenylosuccinate lyase	Pseudogene, Rhs family
Gene Ontology	Galactose metabolic process; lysine biosynthetic process via diaminopimelate	ATP/tRNA binding; Peptidoglycan-based cell wall; Succinate AMP-lyase (fumarate-forming)	Self proteolysis

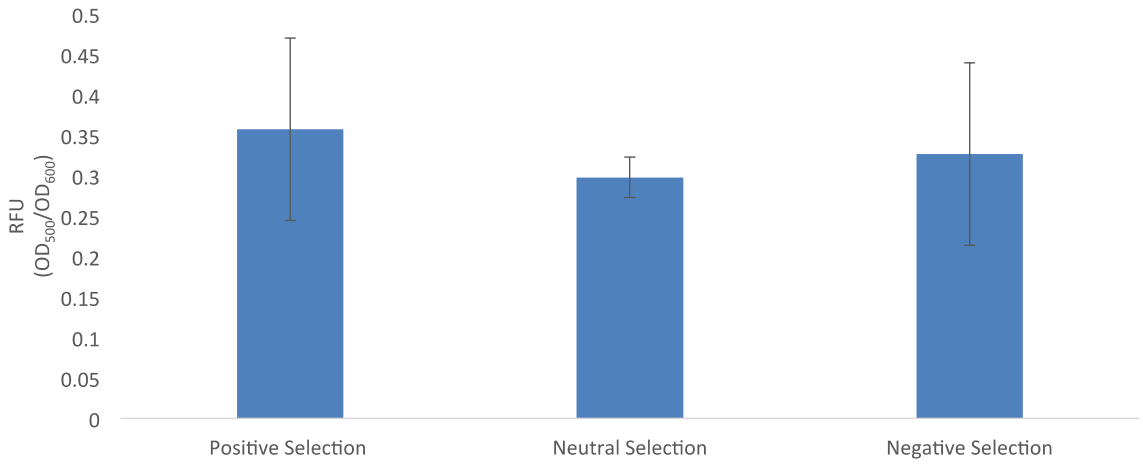


Figure 5-7. L-Arginine analysis of surviving substrains from FITSelect enrichment with (Positive Selection), without (Negative Selection) the inducer molecule OHHL, and common to both enrichments (Neutral Selection). Samples were grown in duplicate and each sample was measured twice for four readings per substrain. Error bars represent one standard deviation.

DISCUSSION

The functioning aspects of FITSelect

The overall goal of the FITSelect gene circuit is to tie the growth rate of a cell to the rate of target product formation. It is assumed that target products will accumulate in intracellular concentration before being exported, so the FITSelect system was tested during construction by adding exogenous target product (i.e., L-arginine in this proof-of-concept case). Reporter genes were used during construction to better understand the function of engineered promoters and gene regulation in the circuit. This proved particularly useful as it enabled us to identify the

original choice of L-arginine responsive promoter, *argF*, did not respond to exogenous L-arginine to regulate expression of the *mCherry* fluorescent reporter gene. However, the *artJ* promoters performed well to this regard, yielding cell growth levels that became inhibited with L-arginine supplementation, with an active response up to around 250 μM exogenous L-arginine. An explanation for the preference of the *artJ* over the *argF* promoter may be that transcriptional control of the L-arginine biosynthesis pathway is most sensitive to feed-back inhibition at the early steps of the reaction pathway [4]. The *artJ* promoter may serve better for the L-arginine responsive circuit in this regard as it controls expression of the ArtJ L-arginine transporter proteins, the point of entry for L-arginine into the cell. On the other hand, the *argF* promoter controls the sixth step in conversion of L-glutamate to L-arginine [8]. By the same analysis, the use of the native *argR* and *carAB* genes have been shown to respond rapidly to the presence of L-arginine [4]. In addition, the *artJ* promoter was found to be selective towards L-arginine (as opposed to L-histidine and L-ornithine), which is a necessary requirement of the FITSelect circuit to ultimately select for high L-arginine producing cells.

The second half of the circuit, leading to *lacZ-ccdB* expression and cell death, was also successful and was shown to be responsive the amount of exogenously added OHHL and the amount of LuxR produced by the L-arginine sensitive part of the circuit. The LacZ-CcdB fusion protein led to cell death yet was also quantifiable by a well-developed β-galactosidase assay. This was a particularly useful aspect of the system that enabled us to quantify FITSelect circuit activity. Thus, together, the FITSelect circuit is responsive to L-arginine concentrations, and cultures grow faster and reach higher densities with increasing L-arginine concentrations. From this perspective, the FITSelect circuit construction and function was successful.

Why the FITSelect circuit failed to enrich for high yield L-arginine producing cells

While the FITSelect circuit functioned will with cells supplemented with exogenous low concentrations of L-arginine, practical enrichment studies failed to a high L-arginine producing substrain in this research (Figure 5-7). There are several reasons for this, and we have proposed solutions that will be examined in future research. First, as shown in Figure 5-3, increases in L-arginine concentration of only 25 μL were required to produce a full response of the L-arginine sensitive promoter. As the concentration of L-arginine was increased to 250 μL, the same response was observed. DNA library fragments giving a large or small response could equally suppress *lacZ-ccdB* gene expression with the current form of the FITSelect circuit. Small increases in L-arginine concentration may not be significant relative to the experimental error associated with measurements. Next, library enrichment may select for DNA fragments that block action of the LacZ-CcdB fusion that leads to cell death. This may be less taxing to the cell than maintaining L-arginine overproduction. It is possible that increasing L-arginine production is a multi-genic trait and that the DNA library did not contain enough genetic diversity (or large enough DNA fragments) to do this and fully take advantage of the circuit. Another possibility is that the enriched fragments increased growth fitness in aspects other than L-arginine production or CcdB resistance, leading to greater growth and fitness in spite of the negative pressure of the FITSelect circuit. The most abundant insert contained the full coding genes *hflD* and *purB*. The first gene product is responsible for resistance against integration of phage DNA [31], and the second catalyzes the formation of fumarate from adenylosuccinate during proline synthesis, regenerate AMP, and aids in acid resistance in *E. coli* [32]. Further work is necessary to elucidate the benefit for cells to select this fragment. These DNA fragments were enriched in the OHHL induced and un-induced cultures, so it is very possible they provide a growth advantage

that provides a small increase in L-arginine production (to satisfy the FITSelect circuit) and then provided additional fitness by another mechanism. Furthermore, we are considering the possibility that high cell densities shield cells from OHHL uptake late in the culture. Cell densities tend to recover towards the end of culture and the FITSelect circuit appears to break-down (seen in Figs. 2, 4, and 6). When this happens, enrichment would likely be skewed away from strong L-arginine producers. Finally, disruption of the *luxR* gene by mutations and recombination of insertion elements [33] has been observed in certain cases. These work to prevent LuxR from inducing expression of the LacZ-CcdB cell death fusion.

The path forward

Modifications of the FITSelect circuit will be made in future research to improve the responsiveness to L-arginine and the stability of the circuit. Genomic integration of FITSelect will improve its stability and remove the cellular energy burden of plasmid replication. It is well known this is particularly effective when expressing foreign DNA [34] and prevents the observed recombination of insertion elements from interrupting the *luxR* gene and overriding the circuit. Genomic integration of FITSelect will also ensure that only a single copy of the circuit exists per cell instead of the 10-15 copies on a p15A-based plasmid. Integration will also help with engineering the L-arginine concentration response, which is of high priority for tuning the FITSelect circuit. Currently, as little as 25 μM of L-arginine is repressing *luxR* expression (Figure 5-3). This means that once the ArgR repressor is activated by L-arginine, it can bind the *arg* boxes of the *artJ* promoter easily. The response needs to be engineered as a more gradual, rather than binary, function through analysis of the binding mechanisms between the activated ArgR repressor and the *arg* box DNA sequences. Additional mutation studies can be

implemented to fine-tune these interactions. Engineering the RBS of the *luxR* and *lacZ-ccdB* mRNAs could be implemented to increase or decrease the amount of LuxR or LacZ-CcdB produced for a given quantity of L-arginine. This could be done to increase the overall response of the circuit and over-ride any competing selection mechanisms. In addition, any residual LuxR in the cell could be reduced by incorporating *ssrA* degradation tags to reduce lag between changes in L-arginine concentration and changes in LuxR abundance [35]. Finally, the role of OHHL uptake in dense cultures must be investigated. It is proposed that cultures undergoing enrichment be grown either in chemostat or with serial dilutions every few doubling times to maintain low OD_{600} values (e.g., between 0.3 and 0.6).

The goal of the project was to create a genetic circuit able to modulate the growth rate of a cell in response to a feedback-controlled metabolite. L-arginine was chosen as the metabolite, both for its industrial and agricultural relevance and the numerous genes controlled by its intracellular concentration. The FITSelect circuit is not limited to just L-arginine. Other promoters within the cell respond to a variety of sugars, amino acids, transcription factors, antibiotics and signaling proteins [36]. The development of SELEX procedures and designer aptamers and riboswitches [37, 38] could also potentially open up the regulation to any small molecule or protein. The target would interact with an engineered aptamer or riboswitch of high specificity and, upon binding, could enable the repression or activation of the cell death gene.

ACKNOWLEDGEMENTS

We acknowledge the contributions of Imen Tanniche, and Manuel Martinez Salgado towards the construction and testing of FITSelect. BGF is supported by the Virginia Tech

Institute for Critical Technology and Advanced Science. We thank Ann M. Stevens for plasmid pJE202 and Warren Ruder for plasmids pKDL071 and pBAD24.

REFERENCES

[1] You, L., Cox, R. S., Weiss, R., Arnold, F. H., Programmed population control by cell-cell communication and regulated killing. *Nature* 2004, *428*, 868-871.

[2] Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., *et al.*, Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotech* 2004, *22*, 1567-1572.

[3] Caldara, M., Minh, P., Bostoen, S., Massant, J., Charlier, D., ArgR-dependent repression of arginine and histidine transport genes in Escherichia coli K-12. *Journal of molecular biology* 2007, *373*, 251.

[4] Zaslaver, A., Mayo, A. E., Rosenberg, R., Bashkin, P., *et al.*, Just-in-time transcription program in metabolic pathways. *Nature genetics* 2004, *36*, 486-491.

[5] Ou, J., Yamada, T., Nagahisa, K., Hirasawa, T., *et al.*, Dynamic change in promoter activation during lysine biosynthesis in Escherichia coli cells. *Mol Biosyst* 2008, *4*, 128-134.

[6] Utagawa, T., Production of Arginine by Fermentation. *The Journal of Nutrition* 2004, *134*, 2854S-2857S.

[7] Caldara, M., Charlier, D., Cunin, R., The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. *Microbiology* 2006, *152*, 3343-3354.

[8] Cunin, R., Glansdorff, N., Piérard, A., Stalon, V., Biosynthesis and metabolism of arginine in bacteria. *Microbiological reviews* 1986, *50*, 314-352.

[9] McKnight, J. R., Satterfield, M. C., Jobgen, W. S., Smith, S. B., *et al.*, Beneficial effects of L-arginine on reducing obesity: potential mechanisms and important implications for human health. *Amino acids* 2010, *39*, 349-357.

[10] Bi'e Tan, X. L., Yin, Y., Wu, Z., Liu, C., *et al.*, Regulatory roles for L-arginine in reducing white adipose tissue. *Frontiers in bioscience: a journal and virtual library* 2012, *17*, 2237.

[11] Shan, Y., Shan, A., Li, J., Zhou, C., Dietary supplementation of arginine and glutamine enhances the growth and intestinal mucosa development of weaned piglets. *Livestock Science*, *150*, 369-373.

[12] Hernandez, A., Hansen, C. F., Mullan, B., Pluske, J., L-arginine supplementation of milk liquid or dry diets fed to pigs after weaning has a positive effect on production in the first three weeks after weaning at 21 days of age. *Animal feed science and technology* 2009, *154*, 102-111.

[13] Kirchgessner, M., Maierhofer, R., Schwarz, F. J., Eidelsburger, U., Effect of feeding protected arginine on food intake, milk yield and growth hormone and amino acid levels in blood plasma of cows during the summer feeding period with grass. *Archiv fur Tierernahrung* 1993, *45*, 57-69.

[14] Al-Daraji, H. J., Salih, A. M., Effect of dietary L-arginine on productive performance of broiler chickens. *Pakistan Journal of Nutrition* 2012, *11*, 252-257.

[15] Fouad, A., El-Senousey, H., Yang, X., Yao, J., Role of Dietary L-arginine in Poultry Production. *International Journal of Poultry Science* 2012, *11*, 718-729.

[16] Ikeda, M., Ohnishi, J., Hayashi, M., Mitsuhashi, S., A genome-based approach to create a minimally mutated Corynebacterium glutamicum strain for efficient L-lysine production. *Journal of Industrial Microbiology and Biotechnology* 2006, *33*, 610-615.

[17] Becker, J., Wittmann, C., Systems and synthetic metabolic engineering for amino acid production – the heartbeat of industrial strain development. *Current Opinion in Biotechnology* 2012, *23*, 718-726.

[18] Kisumi, M., Kato, J., Sugiura, M., Chibata, I., Production of L-arginine by arginine hydroxamate-resistant mutants of Bacillus subtilis. *Applied microbiology* 1971, *22*, 987-991.

[19] Kubota, K., Onoda, T., Kamijo, H., Yoshinaga, F., Okumura, S., Microbial production of L-arginine: Production of L-arginine by mutants of glutamic acid-producing bacteria. *The Journal of General and Applied Microbiology* 1973, *19*, 339-352.

[20] Ikeda, M., Mitsuhashi, S., Tanaka, K., Hayashi, M., Reengineering of a Corynebacterium glutamicum L-arginine and l-Citrulline Producer. *Applied and Environmental Microbiology* 2009, *75*, 1635-1641.

[21] Xu, M., Rao, Z., Xu, H., Lan, C., *et al.*, Enhanced Production of L-arginine by Expression of Vitreoscilla Hemoglobin Using a Novel Expression System in Corynebacterium crenatum. *Applied Biochemistry and Biotechnology* 2011, *163*, 707-719.

[22] Stevens, A., Dolan, K., Greenberg, E., Synergistic binding of the Vibrio fischeri LuxR transcriptional activator domain and RNA polymerase to the lux promoter region. *P Natl Acad Sci USA* 1994, *91*, 12619.

[23] Heckman, K. L., Pease, L. R., Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat. Protocols* 2007, *2*, 924-932.

[24] Guzman, L. M., Belin, D., Carson, M. J., Beckwith, J., Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 1995, *177*, 4121-4130.

[25] Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., Kschischo, M., grofit: fitting biological growth curves with R. *Journal of Statistical Software* 2010, *33*, 1-21.

[26] Nakata, Y., Tang, X., Yokoyama, K., Preparation of Competent Cells for High-Efficiency Plasmid Transformation of Escherichia coli, in: Cowell, I., Austin, C. (Eds.), *cDNA Library Protocols*, Humana Press 1997, pp. 129-137.

[27] Stasyuk, N. Y., Gayda, G. Z., Gonchar, M. V., L-arginine-selective microbial amperometric sensor based on recombinant yeast cells over-producing human liver arginase I. *Sensors and Actuators B: Chemical* 2014, *204*, 515-521.

[28] Salis, H. M., The ribosome binding site calculator. *Methods in enzymology* 2011, *498*, 19-42.

[29] Caldara, M., Dupont, G., Leroy, F., Goldbeter, A., *et al.*, Arginine biosynthesis in Escherichia coli experimental perturbation and mathematical modeling. *Journal of Biological Chemistry* 2008, *283*, 6347-6358.

[30] Freedman, B. G., Zu, T. N. K., Wallace, R., Senger, R. S., Raman spectroscopy characterizes 1-butanol tolerant *E. coli* phenotypes resulting from genomic DNA library enrichment. *Under Review* 2014.

[31] Kihara, A., Akiyama, Y., Ito, K., Revisiting the lysogenization control of bacteriophage lambda. Identification and characterization of a new host component, HflD. *The Journal of biological chemistry* 2001, *276*, 13695-13700.

[32] Sun, Y., Fukamachi, T., Saito, H., Kobayashi, H., ATP Requirement for Acidic Resistance in Escherichia coli. *Journal of Bacteriology* 2011, *193*, 3072-3077.

[33] Saedler, H., Heiß, B., Multiple copies of the insertion-DNA sequences IS1 and IS2 in the chromosome of *E. coli* K-12. *Molecular and General Genetics MGG* 1973, *122*, 267-277.

[34] Yu, D., Ellis, H. M., Lee, E.-C., Jenkins, N. A., Copeland, N. G., An efficient recombination system for chromosome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences* 2000, *97*, 5978-5983.

[35] McGinness, K. E., Baker, T. A., Sauer, R. T., Engineering Controllable Protein Degradation. *Molecular Cell* 2006, *22*, 701-707.

[36] Thieffry, D., Huerta, A. M., Pérez-Rueda, E., Collado-Vides, J., From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *BioEssays* 1998, *20*, 433-440.

[37] Stoltenburg, R., Reinemann, C., Strehlitz, B., SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering* 2007, *24*, 381-403.

[38] Sharma, V., Nomura, Y., Yokobayashi, Y., Engineering Complex Riboswitch Regulation by Dual Genetic Selection. *J Am Chem Soc* 2008, *130*, 16310-16315.

Chapter 6 Conclusions

Library enrichment studies have shown great success in the advancement and improvement in renewable chemical and fuel production. The original goal of the work was to utilize established methods to produce fully function library screens to improve the growth rate and efficiency of *C. cellulolyticum* grown on lignocellulose. After numerous attempts to create a full coverage library of the *C. cellulolyticum* genome, we decided to look into the library assembly procedure itself to reduce the seeming stochastic success rates. When constructing libraries, numerous factors can influence the degree of coverage of the resulting library pools, including:

i. Preparation of the genomic fragments
ii. Reagents used to repair, digest, and modify DNA prior to ligation
iii. Methylation and restriction modifications of the prepared DNA affecting compatibility with the new library host

After many tests, a technique was developed where we could ignore any contribution of the above factors. By using polymerase chain reaction to amplify the initial source of DNA, we could produce a genomic library of the same sequence as the source, but which does not use the same molecules as the host. This technique reduced the effect of methylation as all PCR product will be methylation-free. The use of oligonucleotide primers to initiate each amplification ensures that the terminal ends of the emerging library fragments are identical with no differences in single stranded overhang length or phosphorylation of the termini, often associated with enzymatic digestion or mechanical shearing methods. Perhaps due to the reduction of process variables, the method worked repeatedly and consistently across multiple host genomes. It was highly applicable to amplification of metagenomic DNA, where genomic DNA is difficult to

acquire at quantities needed for traditional genomic libraries. The amplified DNA even looked similar to DNA generated by mechanical shearing methods when viewed on an agarose gel.

Once the method was shown to work, verification was needed to ensure that the DNA generated was an accurate representation of the original genome. High throughput sequencing was used to examine the amplification products and workflows were developed to compare the coverage, distribution, and mutational frequency of the resulting DNA. We found that the initial libraries amplified from an *E. coli* target genome could account for up to 92% of an *E. coli* genetic sequence. Coverage dropped off precipitously for genomes with GC contents 31% and the distribution of the DNA left some regions thousand-fold greater coverage than other regions.

The template design for primers used in degenerate oligonucleotide primed PCR introduced inherent bias to the amplification reaction through the tendency to bind only to a "ATGTTG" sequence. Rational design of the amplification primers allowed replacement of the fixed ATGTTG sequence with a random hexameric N-mer, representing all 4^6 possible binding sites in the genome, but stable enough to form 1kb+ amplification products. Thermodynamic models were then created to further optimize the oligonucleotide design to exclude sequences capable of hair-pinning and secondary structure formation. The improvements enabled up to 97% coverage of the reference genome for *E. coli* and showed a remarkable 300% increase in single base coverage for the recalcitrant *C. acetobutylicum* genome between our experiments, and a 200% increase in representation at 10x coverage compared to the previously published primer design.

The libraries we generated still have much room for improvement. Coverage bias remains the most pressing difference between amplified and unamplified libraries. Further, the cloning steps by which library fragments are incorporated into the expression host can also limit the

diversity of the genomic library before enrichment studies take place. A logical step to monitor this drop-out of representation would be to sequence the library at different stages of construction, up to and including the enrichment process. Not only will this point out bottlenecks in the library cloning process, it will also provide a baseline for future enrichments as changes in the genetic makeup of the expression library could be monitored at a per base level, not just a per gene level reported by currently.

Once libraries were produced which confidently covered the reference, we set out to apply our original enrichment studies. A "proof of concept" in *E. coli* led to rapid selection for fragments from an *E. coli, C. acetobutylicum,* and metagenomic soil library that enabled increased survivability under exposure to 1-butanol. The enriched fragments did match many of the functions identified in previous alcohol tolerance studies and increased cell survivability by up to 100% in some cases. Further, monitoring cell growth of the identified mutants using Raman spectroscopy demonstrated the multiple modes of action by which cells respond to butanol tolerance and the degrees to which they can modify membrane fatty acid content.

Application of a genomic library enrichment in a cellulolytic organism was meant to select for DNA enabling *C. cellulolyticum* with faster cellobiose uptake or more rapid growth rates. Though this process proved most difficult due to the low transformation efficiency of the organism, a small library was developed and enrichment performed to isolate a fragment of the phage minor structural protein which enabled a 250% increase in ethanol concentration but at a cost to growth rate. While difficult to explain from an evolutionary standpoint why the enrichment would select for a slower growing fragment than wild-type, the improvement in ethanol productivity may prove to be useful in future strain engineering efforts.

Our final experiment looked to increase the utility of genomic library enrichments by enabling selection for non-essential traits. Whereas cells exposed to ethanol will select for ethanol tolerant mutants, many traits, such as arginine over-production, cannot be selected for easily. The FITSelect genetic circuit developed here has the potential to develop such a link by tying cell growth rate to the concentration of arginine in culture. While the circuit and all of its regulated components function as expected, the system can only select between "starving" cells with suboptimal arginine levels. Sensitivity and gain require adjustment to differentiate between mutants capable of overproduction.

The methods listed trivialize the production of genomic expression libraries and enable rapid isolation of mutants capable of affecting cell behavior. Monitoring the effect of such changes with Raman spectroscopy can elucidate the physiological mechanism for such behavior. But monitoring the genetic mechanism remains to be performed. Simple reverse-transcription PCR methods or Northern Blots may indicate whether the library insert is being expressed as RNA in the cells. Once verified, we must determine whether the insert produces a regulatory, antisense transcript or a protein coding sequence. Finally, incorporation of the characterized mechanism into a metabolic model, either by balancing flux ratios or addition of metabolic pathways could enable prediction of related or synergistic modifications for further strain improvement.

Chapter 7 Supplementary Materials

Supplementary Appendix A

Thermodynamic design of degenerate oligonucleotide primers for whole genome DNA amplification with reduced bias for use in DNA libraries

Description of methods and analysis

Step 1: Generation of a DNA library with DOP-PCR

Continuous field electrophoresis is the standard method for visualizing smaller DNA fragments in the 40-12000bp range. Whole genomic DNA of bacteria is typically visualized on an agarose gel as a single band of DNA around 12 kb when run in continuous, rather than pulsed field, electrophoresis [1]. The goal of DNA library generation is to turn this single band (i.e., massive piece(s) of DNA) into a series of smaller fragments that, when combined, could reassemble the entire original genome.

The DOP-PCR method involves degenerate primers that bind "randomly" to the genomic DNA, allowing the polymerase to generate a new strand of DNA. During low stringency binding, we expect the most random binding to take place, where each emerging strand could serve as a template for a new primer to bind. During the high stringency cycle, the extra nucleotides in the 5' end of the primer promote annealing to fragments that contain these same nucleotides [2]. Conceptually, this will result in re-amplification of fragments that had been generated previously

with a primer at both ends, and fewer priming events at genomic or internal sites of amplified strands. The resulting DNA "smear," when visualized on a gel, will exhibit a normally distributed shape with maximum intensity between around 3kb and a trailing off around 500 bp and >6000 bp.

Step 2: Physical processing of DNA library for next-generation sequencing

To prepare for Illumina HiSeq next-generation sequencing, the large DNA "smears" are broken down into smaller pieces, averaging 150 bp in length. The small fragments are adenine-tailed, resulting in a single nucleotide overhang to facilitate binding to a specific sequencing adapter with a complementary overhang. The adapter-ligated fragment is amplified with primers specific to the sequencing platform to enable binding to the sequencing flow cell. The fragments are then sequenced, creating a data file where each fragment is given a name, a sequence of up to 150 bp, and a quality score at each base to measure likelihood the base assignment is accurate. This information is provided in a FASTQ file (Fasta + quality).

Step 3: Data Trimming and filtering

Each time the primer binds in DOP-PCR, it leaves extra nucleotides on the newly synthesized fragment. This is referred to as the "barcode," (not to be mistaken for the barcodes added during Step 2) and it must be removed to prevent errors in alignment. Tools like FASTX Barcode Clipper [3] take care of this nicely. Other tools, such as Trimmomatic [4], can trim portions of the reads if the quality dips below a specific threshold. As these steps are preformed, the

following can be collected: (i) the number of sequences that contained the barcode, (ii) the number of reads surviving filtering, and (iii) cleaned sequences ready for analysis. The cleaned file exists as a FASTQ file, with a size slightly smaller than the original one from the sequencing facility.

Step 4: Alignment and re-assembly

If a reference genome is known, such as for *E. coli*, *C. acetobutylicum*, and *C. testosteroni*, an alignment program can be used to assign an address to each short read in the FASTQ file. Bowtie2 is the program used to generate the alignment in this study, and it takes ~10 minutes per sequence running on 4 processing cores [5]. This alignment is stored as a SAM or BAM file. The specifications for SAM/BAM format can be found elsewhere [6], but it contains information related to whether each read from the FASTQ file aligns to the genome, how often it does, its location, orientation, and any mismatches between the read and the aligned region. Once all alignments are made, it is possible to see how much of the reference genome is covered by the sequenced reads and the depth (i.e., number of times each base is covered). Qualimap is a program that automatically generates useful summary statistics of the BAM file, including GC content of the reads that could successfully align. A BED file can also be generated that compiles all the mismatches in the alignment and scores them according to the quality of the mismatch base in the original read sequence and the number of reads that support the mismatch [7].

With the alignment file, random subsamples of defined sizes can be extracted to allow for equal comparisons between sequencing projects. This file was also used to determine where barcoded reads from the FASTX Barcode Clipping program aligned to the reference genome. With this

information, the relationship between the average sequence of the degenerate region of the primer and the actual sequence of the reference can be assessed.

An assembly is a file where the short (~150 bp) reads from a cleaned FASTQ file was combined with no prior knowledge of the reference genome. The nature of the algorithms are beyond the scope of this manuscript, but rely on *de Bruijn* graphs to find overlapping sequences between reads and generate the longest possible continuous sequences (or contigs) [8]. SPAdes is the program used for assembly generation [9]. An accompanying program, QUAST v2.1, will compare the assembly to a reference, identifying mismatches, coverage, overlapping genes, and size distributions [10].

References

[1] Levene, S. D., Zimm, B. H., Separations of open-circular DNA using pulsed-field electrophoresis. *Proceedings of the National Academy of Sciences* 1987, *84*, 4054-4057.

[2] Barbaux, S., Poirier, O., Cambien, F., Use of degenerate oligonucleotide primed PCR (DOP-PCR) for the genotyping of low-concentration DNA samples. *J Mol Med* 2001, *79*, 329-332.

[3] Gordon, A., Hannon, G., Fastx-toolkit: FASTQ/A short-reads pre-processing tools. *(unpublished)* 2010.

[4] Bolger, A. M., Lohse, M., Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, btu170.

[5] Langmead, B., Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012, *9*, 357-359.

[6] Li, H., Handsaker, B., Wysoker, A., Fennell, T., *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, *25*, 2078-2079.

[7] Quinlan, A. R., Hall, I. M., BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, *26*, 841-842.

[8] Compeau, P. E. C., Pevzner, P. A., Tesler, G., How to apply de Bruijn graphs to genome assembly. *Nat Biotech* 2011, *29*, 987-991.

[9] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 2012, *19*, 455-477.

[10] Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013, *29*, 1072-1075.

Supplementary Tables

Table 7-1. All primer designs and results of DOP-PCR amplification. Amplification success was determined by a visual inspection of the resulting gel image of DOP-PCR products. (0) No amplification, (-) faint or uneven amplification, (+) faint but even DNA smear, (++) intense and even DNA smear, and (+++) intense, even, and high molecular weight DNA smear.

Primer Name	Sequence	Total Length (bp)	Amplification Success
Adenyl_Control	GTTTCTTACTCGAGNNNNNNNNNATG	26	++
9NATG	TAGACAATGGGGCATNNNNNNNNNATG	31	++
Adenyl Psos	GTTTCTTGGGGCATNNNNNNNNNATG	26	++
Small_no_tag	NNNNNNNNNATG	12	+
DOP_Midrange	CCGACTCGAGNNNNNNNNNATG	22	+
9N_high_recog_cw	TTAGTAATTCACCGCNNNNNNNNNATG	27	++
Adenyl_Control2	GTTTCTTACTCGAGNNNNNNNNNATG	26	++
DOP9N_psosGib_R	GAAATGGTTTATCTGTTACCCCGTANNNNNNNNNATG	37	0
DOP9N_psosGib_R	CTGGCAAATCATTAAGTGGCGNNNNNNNNNATG	33	0
DOP_9N_NTG_(x4_for_N)	CACCNNNNNNNNNTG	16	-
DOP_LowGC	TATATATATATATATNNNNNN ATG	24	0
DOP_MidGC	CGCTATCGCTATAGCNNNNNNATG	24	+
DOP_HighGC	CGCCGCCGCCGCCGCNNNNNNATG	24	-
DOP_GforN	TAGACAATGGGGCATGGGGGGGGGGATG	27	0
DOP_GN_randaanchor	TAGACAATGGGGCATGGGGGGGGGGNNN	27	+
DOP_6G_3N	TAGACAATGGGGCATGGGGGGNNN	24	+
DOP_6G_6N	TAGACAATGGGGCATGGGGGGNNNNNN	27	-
DOP_9G_6N	TAGACAATGGGGCATGGGGGGGG	30	++

	GNNNNNN		
DOP_9G_2N	TAGACAATGGGGCATGGGGGGGGGNN	26	+
DOP_12G_0N	TAGACAATGGGGCATGGGGGGGGGGGGG	27	-
DOP_MidGC_6G_3N	CGCTATCGCTATAGCGGGGGGNNN	24	+
DOP_MidGC_6G	CGCTATCGCTATAGCGGGGGGATG	24	-
MidGC_6G_6N	CGCTATCGCTATAGCGGGGGGNNNNNN	27	+
MidGC_9G_6N	CGCTATCGCTATAGCGGGGGGGGGNNNNNN	30	+
MidGC_9C_6N	CGCTATCGCTATAGCCCCCCCCCCNNNNNN	30	+
XhoI_MidGC_9G_6N	TGACTCGAGCGCTATCGCTATAGCGGGGGGGGGNNNNNN	39	0
bamHMidGC_9G_6N	TGAGGATCCCGCTATCGCTATAGCGGGGGGGGGNNNNNN	39	0
adenylMidGC_9G_6N	GTTTCCGCTATCGCTATAGCGGGGGGGGGGNNNNNN	35	+
Rational58	TAGTGCCTAGTAAGCGGGGGGGGGGNNNNNN	31	0
Rational60	AGCGTAGTGCCTAGTAGGGGGGGGGNNNNNN	31	+
3bp_tail	ATGTAGACAATGGGGCATGGGGGGGGGNNNNNN	33	+
6bp_tail	ATGATGTAGACAATGGGGCATGGGGGGGGGNNNNNN	36	+
9bp_tail	ATGATGATGTAGACAATGGGGCATGGGGGGGGGNNNNNN	39	+
High_GC_9G_6N	CGCCGCCGCCGCCGCGGGGGGGGGNNNNNN	30	0
DOP_knockdown1	TAGACAATGGGGGGGGGGGGNNNNNN	27	++
DOP_knockdown2	TAGACAATGCATGGGGGGGGGNNNNNN	27	++
DOP_knockdown3	TAGACAGGCATGGGGGGGGGNNNNNN	27	++
DOP_knockdown4	TAGATGGGGCATGGGGGGGGGNNNNNN	27	++
Xho_DOP	TGACTCGAGGGGCATGGGGGGGGGNNNNNN	30	++
BamHI_DOP	TGAGGATCCGGGCATGGGGGGGGGNNNNNN	30	++
Xho_MidGC	TGACTCGAGTATAGCGGGGGGGGGNNNNNN	30	++
Xho_HighGC	TGACTCGAGCGC	30	0

	CGCGGGGGGGGGNNNNNN		
Rand1_9g6n	AGCACTGTCACCGCCGGGGGGGGGNNNNNN	30	0
Rand2_9g6n	GGGCAGAAAATATATGGGGGGGGGNNNNNN	30	++
Rand3_9g6n	ATGCGACGCCTTAAGGGGGGGGGGNNNNNN	30	+++
Rand4_9g6n	GGACCGCTATTGTGAGGGGGGGGGNNNNNN	30	++
Rand5_9g6n	AAGGCAGCAATTAATGGGGGGGGGNNNNNN	30	++
Rand6_9g6n	TCATGCAGGATACCTGGGGGGGGGNNNNNN	30	++
AdenylDOP_9G_6N	GTTTCTTAGGGGCAT GGGGGGGGGNNNNNN	30	++
S1	GATCATGGGGGGNNNNNN	18	0
S2	GATCATCCCCCCNNNNNN	18	-
S3	GATCATGCGCGCNNNNNN	18	+
S4	CTAAAAGGGGGGNNNNNN	18	0
S5	CTAAAACCCCCCNNNNNN	18	+
S6	CTAAAAGCGCGCNNNNNN	19	0
M1	GAGTGGGGTTGGGTTGGGGGGNNNNNN	27	++
M2	TATTTCAAATAATGGGGGGGGNNNNNN	27	++
M3	TATTTCAAATAATGGGCGCGCNNNNNN	27	0
M4	ATATGTTGAACTTATGGGGGGNNNNNN	27	++
M5	ATATGTTGAACTTATCCCCCCNNNNNN	27	++
M6	ATATGTTGAACTTATGCGCGCNNNNNN	27	+
L1	GAGTGGGGTTGGGTTGGGGGGGGGNNNNNN	30	++
L2	TATTTCAAATAATGGGGGGGGGGGNNNNNN	30	+
L4	TATTTCAAATAATCCCCCCCCCCNNNNNN	28	0
L5	ATATGTTGAACTTATCCCCCCCCCCNNNNNN	30	-

Table 7-2. The total numbers of reads and aligned reads for each primer and genome. Also given is the fold amplification of total DNA during DOP-PCR as determined by NanoDrop and Qubit. The *C. acetobutylicum* bacterium contains a megaplasmid (pSOL1). The ratio of pSOL1 plasmid to genome aligned reads is also given.

	***E. coli* NEB 10-beta (50.8% GC content)**			***C. acetobutylicum* ATCC 824 (30.9% GC content)**				***C. testosteroni* ATCC 11996 (61.8% GC content)**		
Primer Name	Total Reads	Aligned Reads	Fold Amplification by NanoDrop (Qubit)	Total Reads	Aligned Reads	Fold Amplification by NanoDrop (Qubit)	Ratio of Plasmid to Genome Aligned Reads[2]	Total Reads	Aligned Reads	Fold Amplification by NanoDrop (Qubit)
None	6,091,805	5,980,679	-	5,345,575	5,255,922	-	1.101	5,327,520	5,122,667	-
Rand3	5,415,632	3,038,415	176 (674)	4,359,573	2,634,015	78 (154)	1.271	4,637,885	3,870,893	25 (1614)
BamHI	5,576,009	3,405,307	384 (697)	6,542,936	1,926,697	159 (888)	0.519	7,207,842	1,929,779	60 (4807)
DOPcontrol	4,828,271	3,467,951	480 (1076)	4,012,005	3,912,443	116 (388)	0.428	4,857,106	1,341,632	66 (10318)
9G 6N	4,914,815	3,241,690	451 (866)	5,939,516	5,515,049	431 (592)	1.178	5,747,454	2,663,791	369 (6423)
9N ATG	6,025,962	2,703,178	356 (1093)	4,948,659	4,103,991	241 (1131)	0.619	7,287,309	277,939[1]	191(5428)
9G 3N	4,726,440	3,915,033	363 (753)	5,002,990	3,956,996	329 (1715)	0.515	5,000,782	3,104,760	350 (5718)
M1	5,023,753	2,583,651	425 (861)	4,467,271	3,429,276	361 (1983)	1.969	6,593,760	2,031,106	226 (4392)
M2	6,775,871	2,569,948	400 (1455)	4,838,183	2,089,751	196 (1161)	1.037	5,452,212	2,212,179	80 (9365)
M4	7,248,864	1,833,948	154 (562)	4,672,657	3,723,436	121 (491)	1.171	6,043,671	1,453,087	181 (11934)
M5	7,362,656	152,076[1]	325 (56)	5,680,019	228,690[1]	124(1983)	0.505	6,930,949	2,233,959	61 (11436)
L1	5,648,871	3,519,483	325 (844)	3,918,919	3,201,648	371 (1983)	1.114	5,303,619	4,566,157	381 (2279)
L2	4,686,253	24,256[1]	265(402)	4,162,614	3,908,591	166 (2431)	1.269	7,637,662	1,238,341	245 (7334)
$MnCl_2$ (L1)	5,214,318	2,980,659	378 (623)	6,843,849	1,354,334	370 (2764)	0.606	7,492,237	2,901,277	520 (9489)
Q5 (L1)	4,234,709	4,111,980	530 (475)	4,238,689	4,108,578	396 (1484)	0.926	4,206,116	3,981,620	614 (11436)
Cold5 (L1)	5,076,290	3,458,747	533 (512)	3,940,638	2,908,461	423 (2111)	1.149	5,506,812	2,926,180	543 (7873)

[1] Amplifications producing fewer than 1M aligned reads were discounted from further analysis

[2] This was calculated as "bases aligned to the plasmid" / "bases aligned to the genome" compared to the proportion of pSOL1 plasmid bases in the reference (4.9%).

Table 7-3. Summary of the number of full genes, partial genes, and number of contigs in genome reconstructions for DNA libraries produced by DOP-PCR.

	***E. coli* NEB 10-beta (50.8% GC content)**			***C. acetobutylicum* ATCC 824 (30.9% GC content)**			***C. testosteroni* ATCC 11996 (61.8% GC content)**		
Primer Name	Full Genes[2]	Partial Genes	Contigs	Full Genes[2]	Partial Genes	Contigs	Full Genes[2]	Partial Genes	Contigs
None	3,512	457	674	3,373	284	377	4,287	754	1,014
Rand3	2,179	1,709	2,621	1,652	1,415	2,213	2,699	2,144	3,271
BamHI	1,612	2,143	3,342	409	2,238	3,646	3,073	1,897	2,745
DOPcontrol	794	1,806	2,491	581	1,387	1,745	1,110	2,752	3,826
9G_6N	1,737	2,080	3,183	901	1,653	2,207	2,584	2,230	3,360
9NATG	1,435	2,051	3,165	547	1,322	1,759	N/A[1]	N/A[1]	N/A[1]
9G_3N	1,103	2,083	2,928	285	824	972	1,377	2,824	4,197
M1	1,494	2,184	3,317	682	1,761	2,214	1,876	2,802	4,351
M2	2,019	1,813	2,978	1,044	2,054	3,003	2,525	2,266	3,459
M4	1,840	1,995	3,238	1,175	1,733	2,646	1,874	2,932	4,777
M5	N/A[1]	N/A[1]	N/A[1]	N/A[1]	N/A[1]	N/A[1]	2,408	2,173	3,437
L1	1,394	2,283	3,313	1,167	2,041	3,109	1,223	3,074	4,622
L2	N/A[1]	N/A[1]	N/A[1]	1,032	1,630	2,227	1,562	2,754	4,128
$MnCl_2$ (L1)	1,055	2,438	3,166	213	1,298	1,725	1,196	3,164	4,798
Q5 (L1)	933	2,447	3,618	819	1,383	1,778	605	1,166	1,464
Cold5 (L1)	1,604	2,065	3,178	745	1,483	1,960	1,583	3,058	4,691
Mset	2,506	1,384	2,127	1,415	1,714	2,551	3,440	1,454	2,200
Lset	1,383	2,298	3,454	1,309	1,789	2,560	1,676	2,738	4,007

[1] Amplifications producing fewer than 1M aligned reads were excluded from analysis
[2] Number of genes in each genome: *E. coli* (4,281), *C. acetobutylicum* (3,885), *C. testosteroni* (5,054).

Supplementary Figures

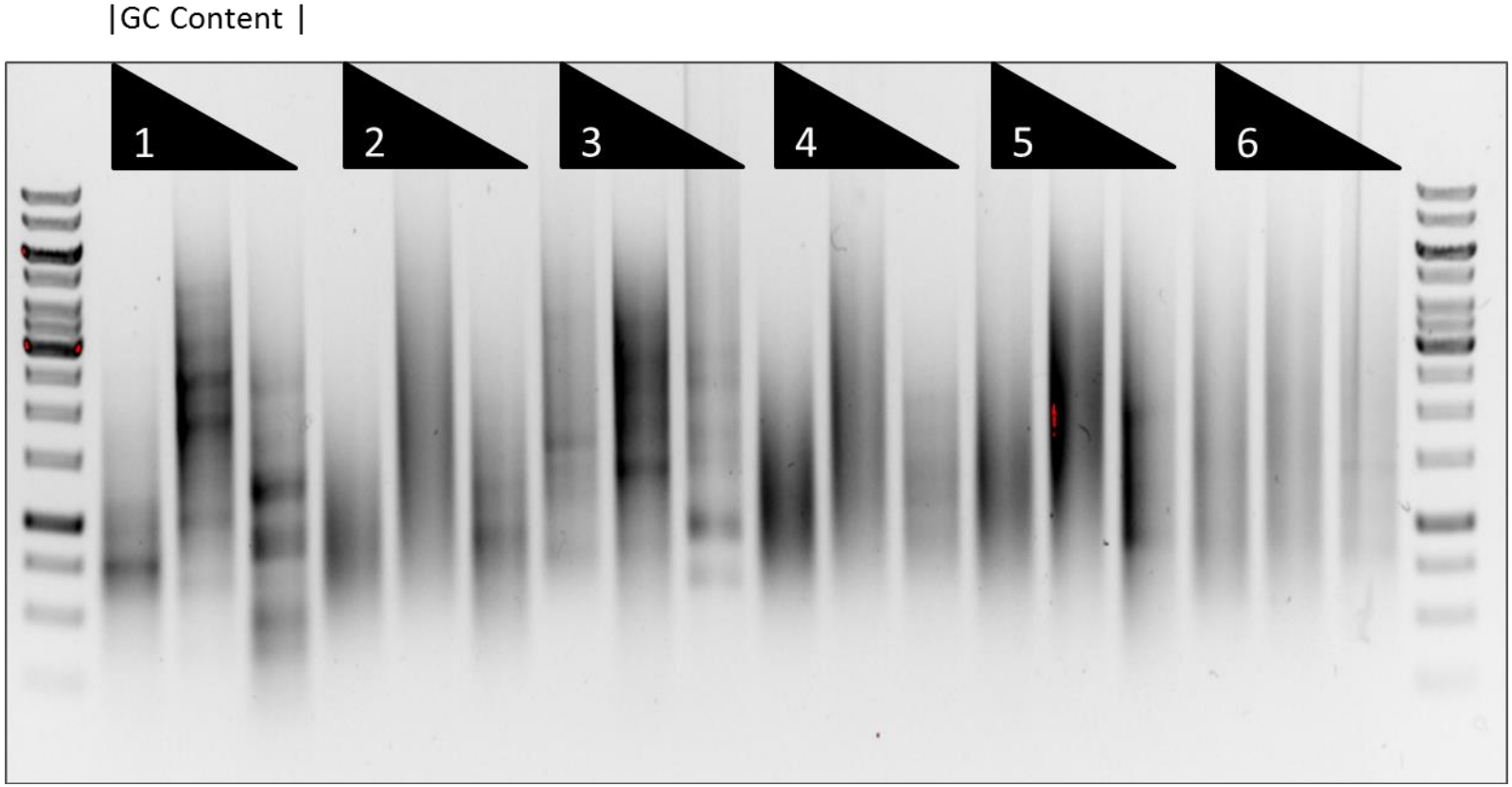


Figure 7-1 Comparison of multiple GC content genomes amplified by DOP under identical conditions. High GC genome: C. testosteroni, Mid GC: E. coli, Low GC: C. acetobutylicum. Primers are (1) DOPcontrol, (2) M1, (3) M5, (4) L1, (5) L2, and (6) Q5 (L1).

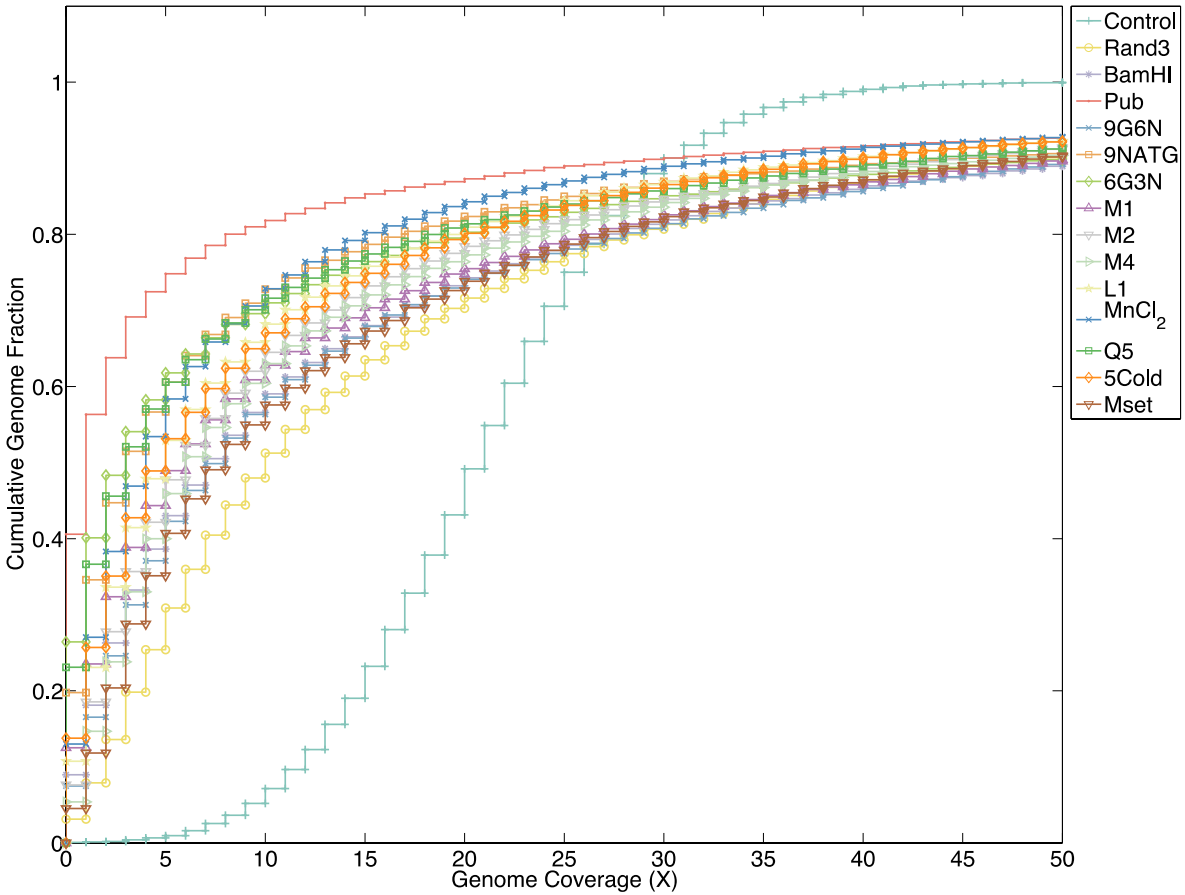


Figure 7-2. Genome coverage and cumulative genome fraction data for E. coli used in K-S statistic calculations.

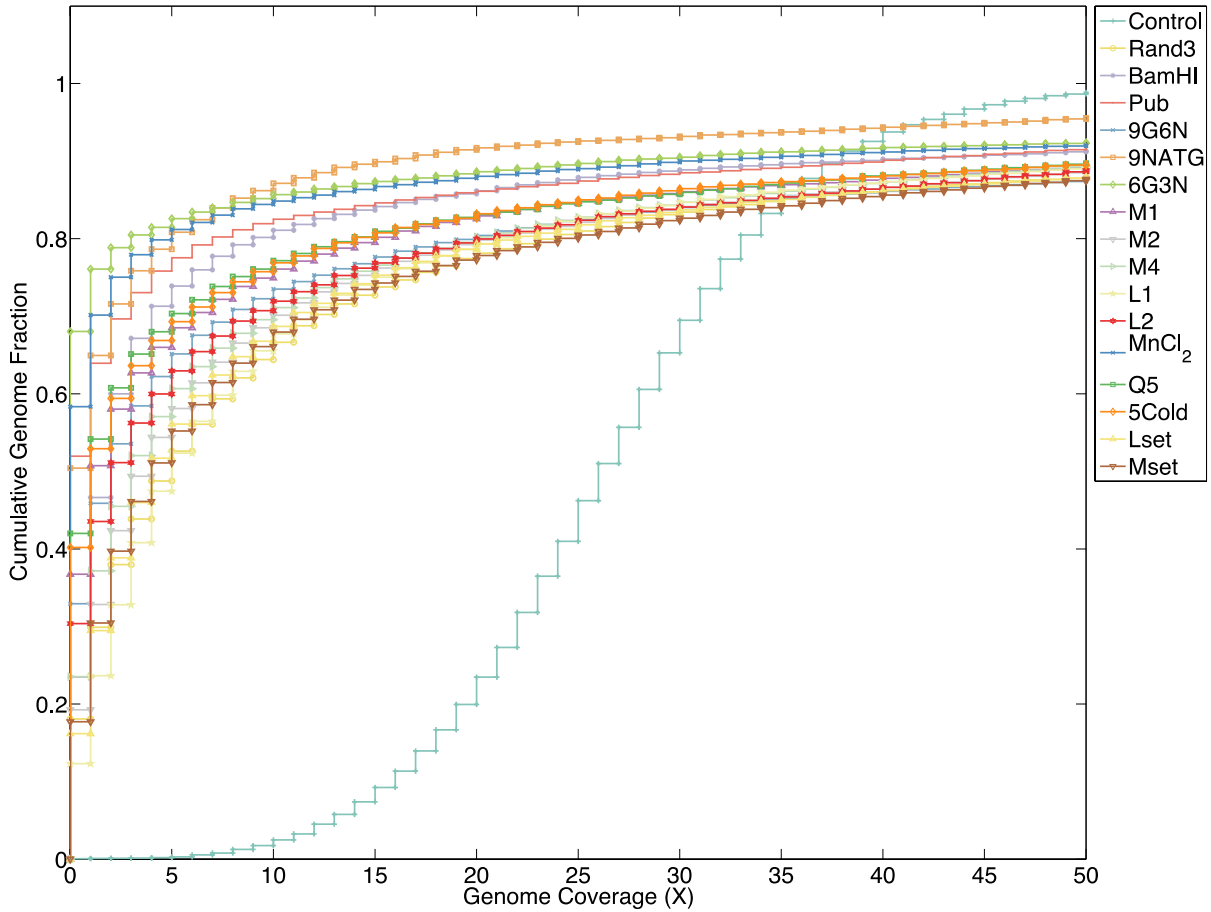


Figure 7-3. Genome coverage and cumulative genome fraction data for C. acetobutylicum used in K-S statistic calculations.

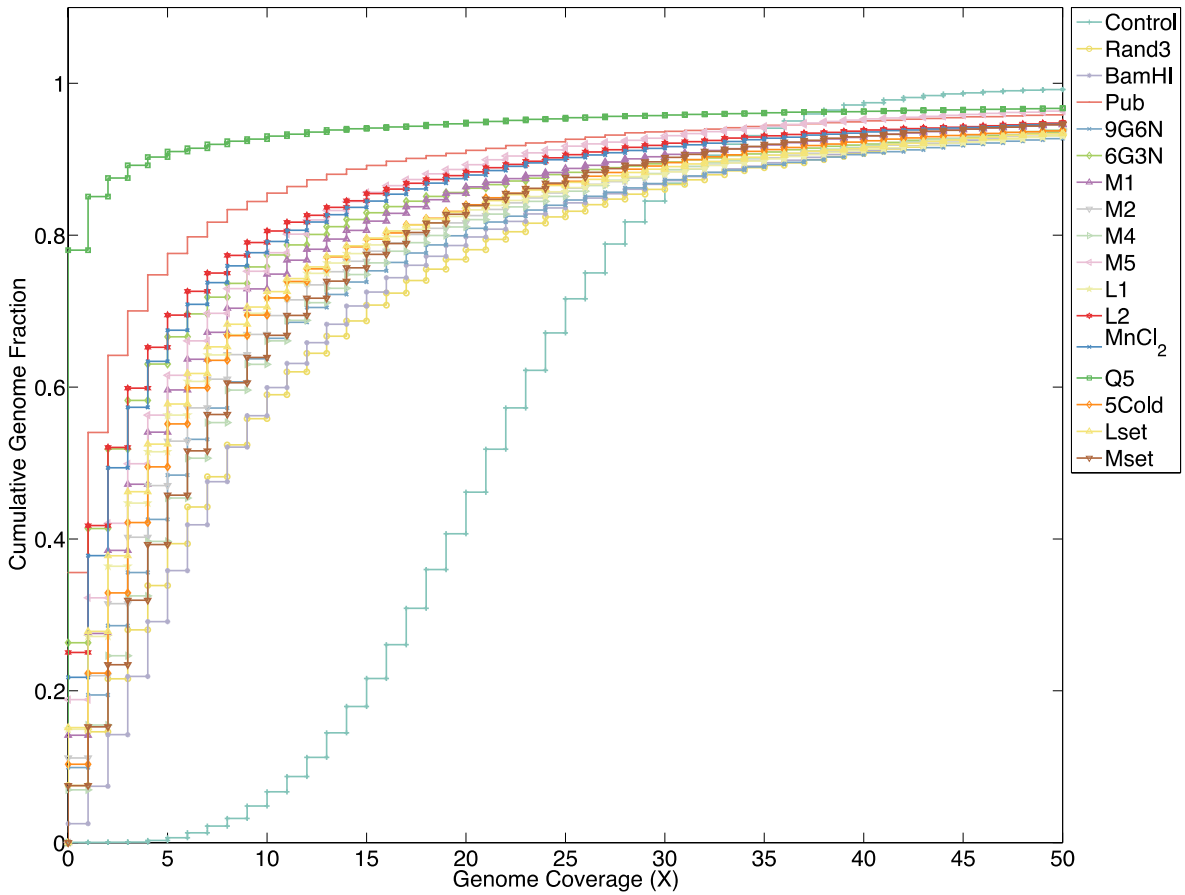


Figure 7-4. Genome coverage and cumulative genome fraction data for C. testosteroni used in K-S statistic calculations.

Supplementary Appendix C

A genomic library to alter the metabolic profile of Clostridium cellulolyticum ATCC 35319

Supplementary Figures

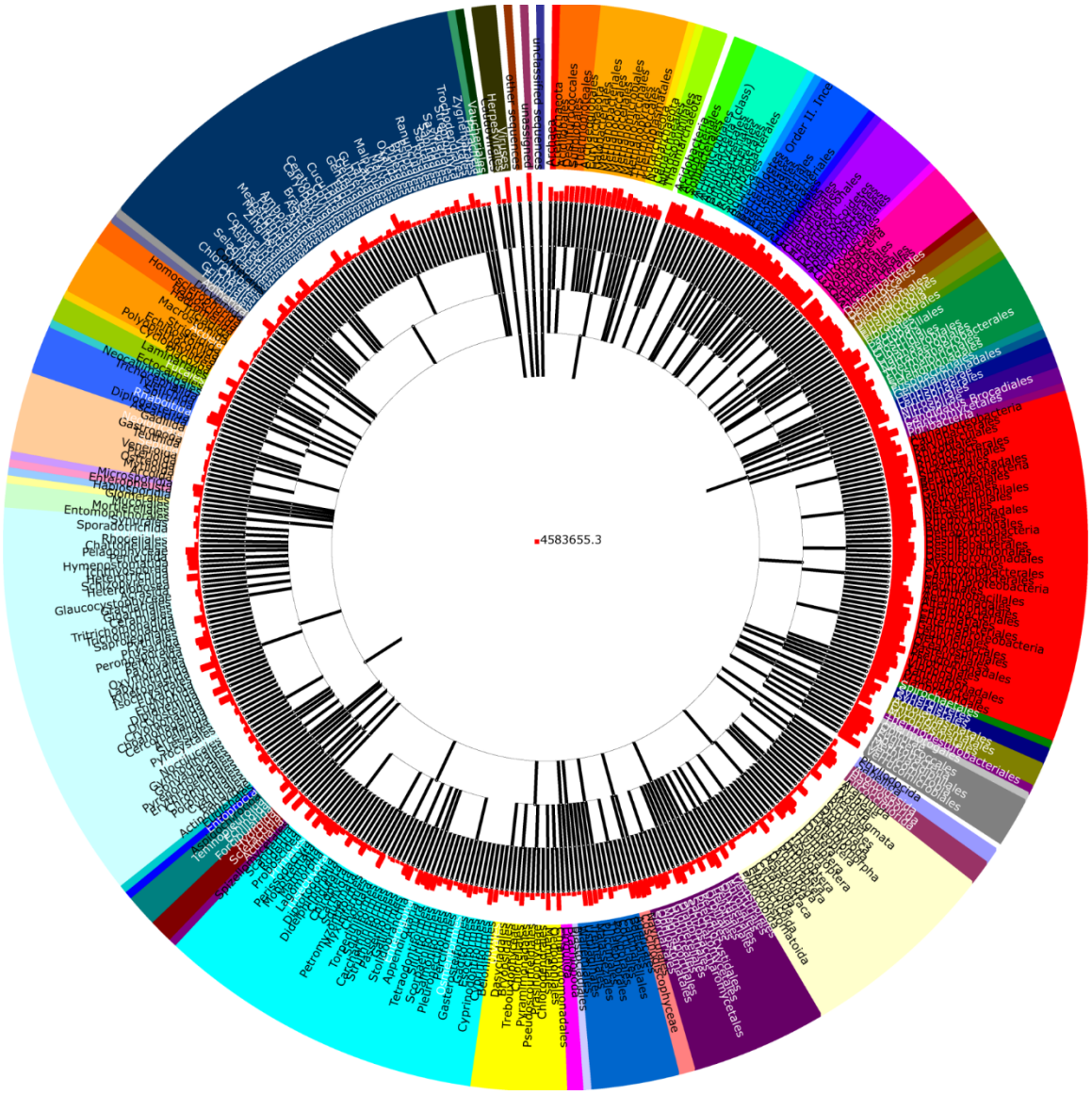


Figure 7-5. Analysis of phylum identified in DOP amplified soil DNA used in metagenomic library. Red bars indicate abundance of reads matching a sample. Black lines indicate lineage trees. Analysis performed using MG-RAST and saved under project number 4583655.3.

Supplementary Appendix B

The Feedback Inhibition of Transcription for Substrain Selection (FITSelect) genetic circuit to couple cell growth rate to non-essential metabolite production

Supplementary Tables

Table 7-4. List of primers and double stranded oligonucleotide Gblocks® from IDT. Underlined sites indicate sequences added for use in restriction digestion

Primer	**Sequence**
LuxR_F	GTTGACAATTAATCATCGAAA TTAATTTTTAAAGTATGGGCAATCAATTG
LuxR_R	GTGGCGTTCGCCATGCGAGGATAAA ATGAAAAACATAAATGCCGAC
Fslux_gblockF	CGTCGGCATTTATGTTTTTCATTTTATCCTCGCATGGCGAAC
Fslux_gblockR	GCAGACCAAAACGATCTCAAAGCACCAATGGGGAGCCG
Osip_AatII_R	TTCTTCGACGTC GTATTATATTCCCCAGAACATCAG
FS_osip_F	ATTGTTGGATCCGGTACCCT GTTGACAATTAATCATCGAAA
FitSel_XbaI	CCCTCTAGAG TTTATATACTAGAGACCTGTAGGATCGTAC
FitSel_EcoRI	GGGGAATTC AAGGTGAGCCAGTGTGACT
LuxR_NdeI	ATTATTCAT ATGAAAAACATAAATGCCGAC
FS_bbone_R	TTTATCCTCGCATGGCGAACG
FS_cher2_F	TCGCCCTTGCTCACCATATG ATCCTCGCATGGCGAACG
gblock_F_bamHI	TGTTGGATCCATCTGACATATG ATCCTCGCATGGCGAAC
Stop_F	CCGGGGATCCTCTAGAGTCG

Stop_R	CGCGAATTCAAAAAGGCCATCCGTCAGGA
Cherry_ndeI	TGGATCAT ATGGTGAGCAAGGGCGAG
Cherry_bamHI	GGATCCGGT TTACTTATACAGCTCGTCCATGC
FS_osip_R	ATTGTTGCTGCAG GTATTATATTCCCCAGAACATCAG
ArtJ_Eco_F	TAGTAGAATTCCGGACAACCCACTAAGTT
ArtJ_WeakRBS	TATCCATATGCTACCTGCTTTTAA CTTATGATTTTTGGCCGTG
T7	TAATACGACTCACTATAGGG
T7_Term	GCTAGTTATTGCTCAGCGG
M1	GAGTGGGGTTGGGTTGGGGGGNNNNNN
M2	TATTTCAAATAATGGGGGGGGNNNNNN
M4	ATATGTTGAACTTATGGGGGGNNNNNN
M5	ATATGTTGAACTTATCCCCCCNNNNNN
Rand3	ATGCGACGCCTTAAGGGGGGGGGGNNNNNN
Double-Stranded DNA	**Sequence**
laczccdB_gblock	TTTATCCTCGCATGGCGAACGCCACTTATTGAATTAAAATTCACTTTATATGTGTAA TTATTCATTTGCAACCCCATTTCACAATTCTTTCTTACAAAGGTGGAGGCAAACCC GTCCGTGTGTGAAAATAATCGAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTT CTGCGTTTATATACTAGAGACCTGTAGGATCGTACAGGTTTACGCAAGAAAATGGT TTGTTATAGTCGAATAAATACTAGAGTCACACAGGAAAGTACTAGATGACCATGAT TACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTT

	ACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGA AGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTATACGTACGGCAG TTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGATGTACA GAGTGATATTATTGACACGCCGGGGCGACGGATGGTGATCCCCCTGGCCAGTGCA CGTCTGCTGTCAGATAAAGTCTCCCGTGAACTTTACCCGGTGGTGCATATCGGGGA TGAAAGCTGGCGCATGATGACCACCGATATGGCCAGTGTGCCGGTCTCCGTTATC GGGGAAGAAGTGGCTGATCTCAGCCACCGCGAAAATGACATCAAAAACGCCATT AACCTGATGTTCTGGGGAATATAATAA
ArtJ_SD_Gblock	CGATGGCCTTTGGCTTGAATTCCGGACAACCCACTAAGTTGTCCGTCTTTTTTATT TCATTTAAATTATTTAATCATGTTTATTGCATATAAATTCACTTGATGGTATTGTTATC CCATGCCGCAGACACGGCCAAAAATCATAAGTTAAAAGCAGGTAGCATATGGTGA GCAAG

Supplementary Figures

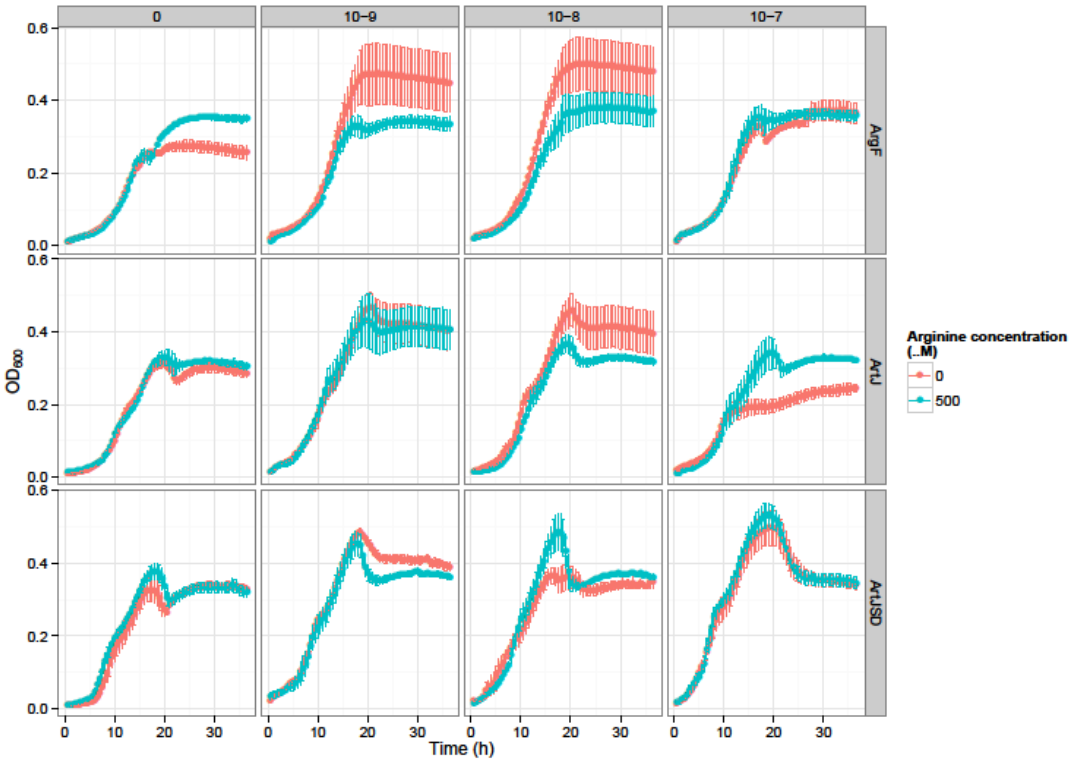


Figure 7-6. Action of ArgF, ArtJ, and ArtJSD mutated promoters controlling expression of luxR in the FITSelect circuit. Cells were grown in minimal media (red) or M9 media with 0.5mM arginine (blue) with OHHL levels of 0, 10-9 ,10-8 , and 10-7M.