INVESTIGATING TRANSCRIPTOMIC RESPONSES TO BIOFUEL STRESS IN *CLOSTRIDIUM ACETOBUTYLICUM*: TRANSCRIPTOME ASSEMBLY AND GENOME ANNOTATION OF A MODEL FERMENTATIVE BACTERIUM.

by

Matthew T. Ralston

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Summer 2014

INVESTIGATING TRANSCRIPTOMIC RESPONSES TO BIOFUEL STRESS IN *CLOSTRIDIUM ACETOBUTYLICUM*: TRANSCRIPTOME ASSEMBLY AND GENOME ANNOTATION OF A MODEL FERMENTATIVE BACTERIUM.

by

Matthew T. Ralston

Approved: _____

Eleftherios T. Papoutsakis, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Errol Lloyd, Ph.D.
Chair of the Department of Computer Science

Approved: _____

Babatunde Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I write this paper with unending thanks for my family, friends, the love, support and encouragement that they give, and the lessons they have taught me. With special thanks for my mother Donna, father Thomas, and sister Allison. With special thanks to my mother; her character and duty towards others is inspirational for this work's focus on sustainability and climate change. With special thanks to my Father, for inspiration through his work ethic, leadership, and open mind. With special thanks to my sister for her support of my education, curiosity, and maturity; I would not be where I am without her encouragement and acceptance. I write this with gratitude to my family for the life they have given me, the sacrifices they have made on my behalf, and most importantly their love. With special thanks for my grandfather George inspiring my pursuit of science and chemistry. With special thanks for my grandmother Winnie for the inspiration of her love and optimism which bring me through each challenge. With special thanks for my uncle Kevin for his curiosity, friendship, optimism, and support. With special thanks for my brother-in-law Matt for his friendship and encouragement. With special thanks for my nieces Violet and Ruby for their love, for the lessons that they teach me, and their infectious energy and optimism. Many thanks for Madeline for her love, open ear, curiosity, support, and encouragement throughout the years. Many thanks to my best friend Andrew for his friendship, curiosity, support, and encouragement. With many thanks to the rest of my wonderful and supportive family and friends for their unconditional love and support.

Many thanks to Karol Miaskiewicz for his consistent and wonderful friendship throughout this project. Many thanks to all of the members of the Bioinformatics program. Particularly, I'd like to thank Erin Crowgey for her mentorship and support of my efforts to learn NGS bioinformatic analyses and Ryan Moore for tossing ideas

around together. I'd like to thank Shawn Polson for his open ear and perspective. I'd also like to thank Dr. Wu for her support and encouragement throughout the program and the many members of the Wu group for their support as well. Many thanks to Bruce, Olga, and Summer for their work in the Sequencing and Genotyping Center in support of this project.

Many thanks for my mentors, Drs. Keerthi Venkataramanan and Terry Papoutsakis, for their mentorship, support, critique, and understanding throughout this project. The success that we've seen in this effort I owe to Keerthi's guidance and training. Many thanks for each of the the many members of the Papoutsakis lab for their friendship, support, and critique. I am very grateful for the unity of this group and for how much I enjoy coming to the lab each day. For the hard work, sleepless nights, early autoclave cycles, shared spaces, and plenty of reasons to celebrate I am grateful for each member.

Ad maiorem dei gloriam.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Chapter 1

METHODS

1.1 Culture

Wild type *Clostridium acetobutylicum* ATCC 824 was cultured anaerobically in 4L New Brunswick Scientific BioFlo 310 bioreactors at 37 °C, pH ¿5.0, 200 mL min^{-1} N_2 and 200rpm agitation in a defined *Clostridia* growth medium, as described previously (Venkataramanan 2013). When the cultures were grown to A_{600}=1, the N_2 flow rate was decreased to 50 mL min^{-1} and cultures were either stressed to a final concentration of 60 mM *n*-butanol, 40 mM potassium butyrate, or left unstressed. 15 mL samples were acquired at 15, 75, 150, and 270 minutes after treatment and synchronization. Samples were centrifuged at 8,000rpm, 4 °C for 20 minutes. After discarding the supernatant, cell pellets were then immediately frozen at −85 °C.

1.2 RNA preparation

RNA was extracted by first washing the cell pellets in 1mL of RNase-free SET buffer (25% sucrose, 50 mM EDTA [pH 8.0], 50 mM Tris-HCl [pH 8.0]) before resuspending cells in a 220 mL solution of RNase-free SET buffer containing 4.55 U mL^{-1} proteinase K and 20 mg mL^{-1} lysozyme and incubating for 6 minutes. Resuspended cells were vortexed with 40mg of RNase-free glass beads (≤106 µm) at maximum speed and room temperature for 4 minutes. Each sample was mixed immediately with 1 mL of ice-cold QIAzol (Qiagen, Valencia, CA, USA) and then 200 µL of ice-cold chloroform, mixing well. After a 3 minute room temperature incubation, samples were centrifuged at 11,000rpm and 4 °C for 15 minutes. The aqueous phase was then mixed with 1.3 mL of ice-cold ethanol before transfering to a miRNeasy Mini spin-column (Qiagen, Valencia, CA, USA) and centrifuging at 11,000rpm and 4 °C for 15 seconds.

Next, 700 µL of RWT buffer was added to the column, before centrifuging at 11,000rpm and 4 °C for 15 seconds, discarding the collection tube and transferring the column to a fresh collection tube. The column was washed twice with 500 µL of RPE buffer before centrifuging at 11,000rpm and 4 degreeCelsius for 15 seconds each. The membrane was then dried with an additional centrifugation step at 11,000rpm and 4 degreeCelsius for 1 minute. The RNA was eluted twice by incubating with 50 µL of nuclease-free water for 1 minute and eluting for 1 minute at 11,000rpm and 4 degreeCelsius.

After quantification on a Nanodrop ND-1000, samples were then precipitated in 0.3M sodium acetate and 75% ethanol overnight, centrifuged at 14,000 rpm for 30 minutes, washed twice with 400 µL ice-cold 70% ethanol, and rehydrated in 50 µL RNase-free water. Next, samples were treated with the Turbo DNA-free kit (Ambion, Austin, TX, USA). 5 µL of 10X Turbo DNase buffer and 1 µL of Turbo DNase ($2\mathrm{U}\mu\mathrm{L}^{-1}$) were added to each sample before incubating at 37 degreeCelsius for 30 minutes. Next, 5 µL of DNase inactivation reagent were added to each sample, mixing occasionally for 5 minutes. The samples were then centrifuged at 10,000rpm and 4 degreeCelsius for 90 seconds, precipitating the DNase.

Samples were then precipitated, washed twice more with 70% ethanol, and resuspended in 20 µL of nuclease-free water, requantified, and aliquoted for quality analysis with the BioAnalyzer platform (Agilent, Wilmington, DE, USA), and 10 µg aliquots in 10 µL samples were stored at −85 °C.

1.3 RNA enrichment, RNA-seq library preparation, and Sequencing

Ribosomal RNA was removed with the MicrobExpress kit (Ambion, Austin, TX, USA) according to their protocol. Briefly, beads were prepared by taking 50 µL for each sample, washing with an equal volume (50 µL) of water capturing for 5 minutes on a MagnaSphere (Promega, Madison, WI, USA) magnetic stand and aspirating. Subsequently, the beads were resuspended in an equal volume (50 µL each) of binding buffer and capturing as above. The beads were then resuspended in an equal volume

(50 µL each) of binding buffer and warmed to 37 °C. Next, 200 µL of binding buffer was added to each 10 µg RNA aliquot with 4 µL of capture oligo mix. The mixture was warmed to 70 °C for 10 minutes, then cooled to 37 °C for 15 minutes. Next, the rRNA was captured by mixing 50 µL of beads with each sample, incubating for 15 minutes at 37 °C, and capturing as above. The enriched RNA was transferred to a fresh 1.5 mL tube. The beads were then washed with 100 µL of pre-warmed (37 °C) wash solution, incubating on the magnetic stand for 5 minutes, and adding the wash solution to the enriched RNA. The samples were then ethanol precipitated at 20 °C overnight with 35 µL of 3 M Sodium Acetate, 5 mg mL^{-1} Glycogen, and 1175 µL of chilled 100% ethanol. The samples were washed twice with 70% ethanol and resuspended in 25 µL. The samples were enriched further by repeating the MicrobExpress treatment. Small 10-100 ng aliquots were analyzed at each step with the BioAnalyzer to monitor enrichment.

Selected samples were enriched further with Terminator 5'-phosphate dependent exonuclease kit (Epicentre, Madison, WI, USA). Terminator Exonuclease 1 µL (1U µL^{-1}) was added with 2 µL 10X Buffer A to each RNA sample. The reaction was run in a thermocycler for 60 minutes at 30 °C. The reaction was terminated with the addition of 1 µL of 100 mM EDTA and ···· Tris HCl at pH 8.0. The samples were then purified by ethanol precipitation (0.3 M Sodium Acetate and 75% ethanol) with two 70% ethanol washes, as above. Enriched RNA was quantified as above and assessed for quality with the BioAnalyzer platform (Agilent, Wilmington, DE, USA). High quality samples were used to prepare RNA-seq libraries with the ScriptSeq v2 library preparation kit and indexed PCR primers (Epicentre, Madison, WI, USA). Briefly, 1 µL of fragmentation solution and 2 µL of cDNA synthesis primer was added to 50 ng of RNA and the solution was fragmented for 5 minutes at 85 °C in a ····· thermocycler. To each reaction, 0.5 mM of Dithiothreitol, 3 µL of cDNA synthesis premix, 0.5 µL StarScript Reverse Transcriptase. is added to each sample and run with the following cycle: 5 minutes at 25 °C, 20 minutes at 42 °C. After cooling each reaction to 37 °C, 1 µL of finishing solution was added, incubating for 10 minutes. The RNA is degraded by

fragmenting further for 3 minutes at 95 °C, cooling to 25 °C. The first strand cDNA is di-tagged by adding 7.5 µL of terminal tagging premix and 0.5 µL of DNA polymerase. The terminal tagging reaction is run at 25 °C for 15 minutes and 95 °C for 3 minutes. The di-tagget cDNA is then purified with the AMPure XP bead system (Beckmann Coulter, Brea, CA, USA). First, the library is mixed with 45 µL of homogenous bead mixture. After thorough mixing, each solution is transferred to a 1.5 mL tube and the library is captured with the magnetic stand and the supernatant aspirated. Each library is then washed twice with 200 µL of 80% ethanol. After resuspending in 24.5 µL of nuclease-free water, the beads are captured and each library is transferred to a new 200 µL microfuge tube. Adapters are added to the di-tagged cDNA during PCR by adding 25 µL FailSafe Premix E, 1 µL forward primer, 1 µL of ScriptSeq v2 indexed reverse PCR primer, 0.5 µL of FailSafe Polymerase. The PCR conditions are as follows: cycles of 30 seconds of 95 °C, 30 seconds of 55 °C, and 3 minutes of 68 °C. After 12 cycles, the reaction terminates with a 7 minute incubation at 68 °C before purifying the library with the AMPure system, as above. Libraries were multiplexed and sequenced for 101 cycles over two lanes of an Illumina HiSeq 2500 at the University of Delaware Sequencing and Genotyping Center (Newark, DE, USA).

1.4 Data Analysis

Paired-end sequencing resulted in pairs of 100bp reads which are deposited in the Sequence Read Archive ( ). Summary statistics for the libraries are shown in table/appendix ( ). The basic bioinformatic processing pipeline is described on Github. In brief, after removing adapters and low quality bases with Trimmomatic, reads were first mapped to the ribosomal RNA sequences before mapping to the *C. acetobutylicum* genome both with Bowtie 2.1.0 [parameters:]. The resulting alignment was then cleaned, validated, sorted, and indexed with SAMtools and Picard. Unmapped and duplicate reads were flagged for omission from further analysis. Reference transcriptome assembly was performed with the Cufflinks suite (v. 2.2.0) or Trinity. The

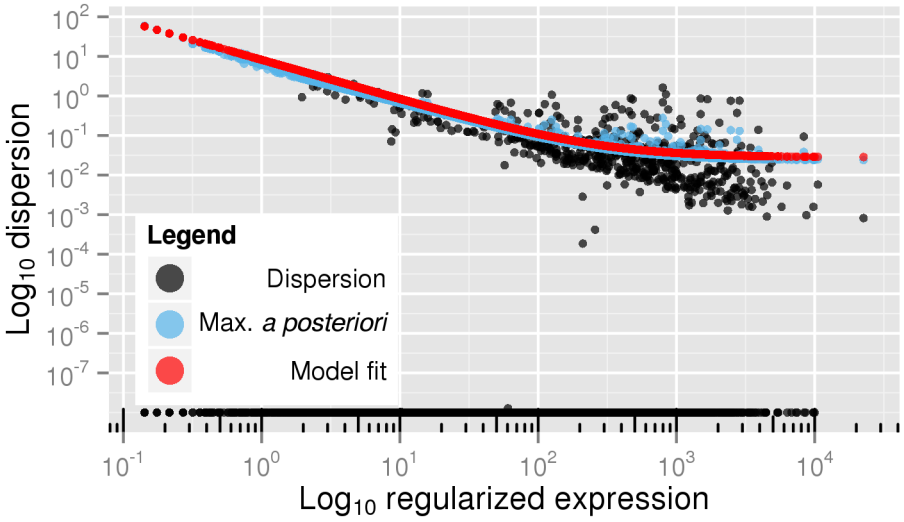
coverage vectors for each strand was calculated by assembling proper pairs into fragments in a BED file, and calculating coverage with BEDtools. Transcriptional start sites were identified with with the peak-finding algorithms TSSi. The fully merged transcriptome assembly were used to determine digital gene expression for visualization and statistical analyses in R and Circos. PCA and clustering was performed in R with and promoter predictions were performed with MEME, and RSAT.

1.5 Molecular Methods

Fold changes were confirmed for randomly selected genes by qRT-PCR. Transcriptional start sites were verified by 5′RACE for randomly selected genes. Small RNAs were confirmed by RT-PCR and Northern blot.

Chapter 2

RESULTS



Chapter 3

DISCUSSION

Chapter 4

CONCLUSIONS

BIBLIOGRAPHY

[1] Keith V Alsaker, Thomas R Spitzer, and Eleftherios T Papoutsakis. Transcriptional analysis of spo0a overexpression in clostridium acetobutylicum and its effect on the cell's response to butanol stress. *Journal of bacteriology*, 186(7):1959–1971, 2004.

[2] S Andrews. Fastqc: A quality control tool for high throughput sequence data. *Reference Source*, 2010.

[3] Dominik Antoni, Vladimir V Zverlov, and Wolfgang H Schwarz. Biofuels from microbes. *Applied microbiology and biotechnology*, 77(1):23–35, 2007.

[4] Shota Atsumi, Anthony F Cann, Michael R Connor, Claire R Shen, Kevin M Smith, Mark P Brynildsen, Katherine JY Chou, Taizo Hanai, and James C Liao. Metabolic engineering of¡ i¿ escherichia coli¡/i¿ for 1-butanol production. *Metabolic engineering*, 10(6):305–311, 2008.

[5] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl 2):W202–W208, 2009.

[6] Jacob R Borden and Eleftherios Terry Papoutsakis. Dynamics of genomic-library enrichment and identification of solvent tolerance genes for clostridium acetobutylicum. *Applied and environmental microbiology*, 73(9):3061–3068, 2007.

[7] Yili Chen, Dinesh C Indurthi, Shawn W Jones, and Eleftherios T Papoutsakis. Small rnas in the genus clostridium. *MBio*, 2(1):e00340–10, 2011.

[8] Patrik Dhaeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[9] A Gordon and GJ Hannon. Fastx-toolkit. *FASTQ short-reads pre-processing tools*, 2010.

[10] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic modelsa review. *Biosystems*, 96(1):86–103, 2009.

[11] Michael Hecker, Wolfgang Schumann, and Uwe Vlker. Heatshock and general stress response in bacillus subtilis. *Molecular microbiology*, 19(3):417–428, 1996.

[12] Eric CH Ho, Michael E Donaldson, and Barry J Saville. *Detection of antisense RNA transcripts by strand-specific RT-PCR*, pages 125–138. Springer, 2010.

[13] Jay D Keasling, Abraham Mendoza, and Phil S Baran. Synthesis: A constructive debate. *Nature*, 492(7428):188–189, 2012.

[14] Donghyuk Kim, Jay Sung-Joong Hong, Yu Qiu, Harish Nagarajan, Joo-Hyun Seo, Byung-Kwan Cho, Shih-Feng Tsai, and Bernhard Palsson. Comparative analysis of regulatory elements between escherichia coli and klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS genetics*, 8(8):e1002867, 2012.

[15] Clemens Kreutz, JS Gehring, D Lang, Ralf Reski, Jens Timmer, and Stefan A Rensing. Tssian r package for transcription start site identification from 5 mrna tag data. *Bioinformatics*, 28(12):1641–1642, 2012.

[16] Ethan I Lan and James C Liao. Metabolic engineering of cyanobacteria for 1-butanol production from carbon dioxide. *Metabolic engineering*, 13(4):353–363, 2011.

[17] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[18] Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. Comprehensive comparative analysis of strand-specific rna sequencing methods. *Nature methods*, 7(9):709–715, 2010.

[19] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[20] Ronny Lorenz, Stephan HF Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[21] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bhlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

[22] Florian Markowetz, Dennis Kostka, Olga G Troyanskaya, and Rainer Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–i312, 2007.

[23] G Najoshi. Sickle-a windowed adaptive trimming tool for fastq files using quality.

[24] Sergios A Nicolaou, Stefan M Gaida, and Eleftherios T Papoutsakis. A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: from biofuels and chemicals, to biocatalysis and bioremediation. *Metabolic engineering*, 12(4):307–331, 2010.

[25] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98, 2011.

[26] Eleftherios T Papoutsakis. Engineering solventogenic clostridia. *Current opinion in biotechnology*, 19(5):420–429, 2008.

[27] Christophe Pichon and Brice Felden. Small rna gene identification and mrna target predictions in bacteria. *Bioinformatics*, 24(24):2807–2813, 2008.

[28] Yosef Prat, Menachem Fromer, Nathan Linial, and Michal Linial. Recovering key biological constituents through sparse representation of gene expression. *Bioinformatics*, 27(5):655–661, 2011.

[29] Juan L Ramos, Estrella Duque, Mara-Trinidad Gallegos, Patricia Godoy, Mara Isabel Ramos-Gonzlez, Antonia Rojas, Wilson Tern, and Ana Segura. Mechanisms of solvent tolerance in gram-negative bacteria. *Annual Reviews in Microbiology*, 56(1):743–768, 2002.

[30] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[31] Tobias Sahr, Christophe Rusniok, Delphine Dervins-Ravault, Odile Sismeiro, Jean-Yves Coppee, and Carmen Buchrieser. Deep sequencing defines the transcriptional map of l. pneumophila and identifies growth phase-dependent regulated ncrnas implicated in virulence. *RNA Biol*, 9(4):503–519, 2012.

[32] Yogita Sardessai and Saroj Bhosle. Tolerance of bacteria to organic solvents. *Research in Microbiology*, 153(5):263–268, 2002.

[33] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

[34] Olga A Soutourina, Marc Monot, Pierre Boudry, Laure Saujet, Christophe Pichon, Odile Sismeiro, Ekaterina Semenova, Konstantin Severinov, Chantal Le Bouguenec, and Jean-Yves Coppe. Genome-wide identification of regulatory rnas in the human pathogen clostridium difficile. *PLoS genetics*, 9(5):e1003493, 2013.

[35] Morgane Thomas-Chollier, Matthieu Defrance, Alejandra Medina-Rivera, Olivier Sand, Carl Herrmann, Denis Thieffry, and Jacques van Helden. Rsat 2011: regulatory sequence analysis tools. *Nucleic acids research*, 39(suppl 2):W86–W91, 2011.

[36] Christopher A Tomas, Jeffrey Beamish, and Eleftherios T Papoutsakis. Transcriptional analysis of butanol stress and tolerance in clostridium acetobutylicum. *Journal of bacteriology*, 186(7):2006–2018, 2004.

[37] Bryan P Tracy, Shawn W Jones, Alan G Fast, Dinesh C Indurthi, and Eleftherios T Papoutsakis. Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. *Current opinion in biotechnology*, 23(3):364–381, 2012.

[38] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[39] Keerthi P Venkataramanan, Shawn W Jones, Kevin P McCormick, Sridhara G Kunjeti, Matthew T Ralston, Blake C Meyers, and Eleftherios T Papoutsakis. The clostridium small rnome that responds to stress: the paradigm and importance of toxic metabolite stress in c. acetobutylicum. *BMC genomics*, 14(1):849, 2013.

[40] Alec W. Picard, 2009.

[41] E Gerhart H Wagner. Kill the messenger: bacterial antisense rna promotes mrna decay. *Nature structural and molecular biology*, 16(8):804–806, 2009.

[42] Qinghua Wang, Keerthi Prasad Venkataramanan, Hongzhan Huang, Eleftherios T Papoutsakis, and Cathy H Wu. Transcription factors and genetic circuits orchestrating the complex, multilayered response of clostridium acetobutylicum to butanol and butyrate stress. *BMC systems biology*, 7(1):120, 2013.

[43] Ka Yee Yeung and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[44] Yong Zhang, Tao Liu, Clifford A Meyer, Jrme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, and Wei Li. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

[45] Kyle A Zingaro and Eleftherios Terry Papoutsakis. Toward a semisynthetic stress response system to engineer microbial solvent tolerance. *Mbio*, 3(5):e00308–12, 2012.

[46] Kyle A Zingaro and Eleftherios Terry Papoutsakis. Groesl overexpression imparts¡ i¿ escherichia coli¡/i¿ tolerance to¡ i¿ i¡/i¿-,¡ i¿ n¡/i¿-, and 2-butanol, 1, 2, 4-butanetriol and ethanol with complex and unpredictable patterns. *Metabolic engineering*, 15:196–205, 2013.