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

Chapter

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Chapter 1
INTRODUCTION

1.1 Motivation

Increases in global CO2 levels, sea level, temperature, and acidification are tipping climate models toward disaster.(UN AR5). Few solutions exist for what has been described as the "...issue that will define the contours of this country more dramatically than any other." (OBAMA) A chief issue with the global CO2 equation is the lack of systems which consume the greenhouse gas. A renewable chemicals industry has been suggested to restore balance to our climate system. No economically effective system exists to convert CO2 into high-quality fuels and chemicals, though the magnitude of the first-generation biofuels industry has demonstrated the scalability of bioconversions.

Leading scientists and engineers (Liao, S.y.Lee, Papoutsakis) recognize Clostridum acetobutylicum as a potential platform organism for a 'biorefinery'. This organism produces an advanced biofuel called butanol, a gasoline replacement being investigated by Gevo, DuPont, and BP. Over the last (X) years, genetic tools have been developed to optimize C. acetobutylicum productivity. Metabolic engineering techniques are being used to increase butanol yield and its already impressive feedstock flexibility. However, successful strains also require biosystems engineering to increase robustness and stress tolerance. Limited knowledge of the stress response systems in C. acetobutylicum is a barrier to engineering robust strains for renewable fuel production.

1.2 Approach

To address this issue, I investigated the response of the transcriptome to biofuel and metabolite stress using Next-Generation Sequencing. I used a mixture of laboratory and informatic approaches to understand the restructuring of the transcriptome under small-molecule stress. RNA-sequencing allowed both the assembly and categorization of transcripts and small RNAs. To make these data accessible to metabolic and biosystems engineers, I constructed interactive visualizations for these results including MA-plots, significance criteria, clustering results, and coverage vectors.

1.3 Document Overview

This document describes the laboratory and informatic procedures used to generate a comprehensive view of a state of the transcriptome and how to compare these states to generate hypotheses. Here I describe the first reported transcriptome assembly and annotation in Clostridium acetobutylicum, a model gram-positive anaerobe and platform organism for biofuel production. Additional results include novel transcripts, CDSes (possible new systems), and TFBSes. I discuss hypotheses that could explain the interesting observations. I discuss the impact of this work and how it should be used to increase productivity in microbial fermentations.

Chapter 2

BACKGROUND

2.1 History of the Involvement of Clostridium acetobutylicum in Industrial Processes

Industrialization and suburbanization have led to dramatic increases in CO2 levels over the 20th century.(REVIEWS) Current climate models suggest that increases in ocean levels, temperature, glacial melt/polar ice reserves, variations in weather patterns are attributable to the rise in CO2 levels.(REVIEWS) Meanwhile, petrochemical exploration suggests a dwindling amount of new prospects. (CITATIONS) There is both an economic and moral imperative to developing renewable strategies for fuel and chemical production.

Our modern and behemoth petrochemical system is limited by finite petroleum reserves. The renewables industry requires processes that utilize abundant and economical resources. Good feedstocks have CO2 as the ultimate Carbon source in the supply chain. The best microorganisms for these conversions can consume a diverse range of substrates and convert to product at high yield. An excellent biofuel would be renewable, economical, and compatible with existing combustion technology and infrastructure. Government agencies (EPA,DOE) have encouraged the growth of the biofuels industry with grants, tax credits, and other financial incentives. The history of recent biofuels development is briefly explored.

First generation biofuels such as ethanol and conventional biodiesel use sugars and fats and input, thus affecting the supply of food. Ethanol has additional issues concerning its hygroscopicity and energy content. It decreases combustion efficiency (source?) and has a negative effect on many conventional systems when used in high concentrations. There have been additional concerns raised about the net energy benefits of ethanol (PNAS study, others, DOE?). The first generation of biofuels compete with consumers for supply and fail to use more abundant sources of carbon. Second generation biofuels focus on the use of non-food crops and waste biomass. Many fuel molecules have been considered, but lignocellulosic ethanol is the classic example. Producing fuels from non-food crops and industrial byproducts has the advantage of cheap feedstock. However, the processes required to digest lignin and cellulosic carbon sources tend to involve expensive pretreatments (examples? sources?). So far, US companies have failed to meet EPA/DOE second generation fuel production requirements by X gallons. An inexpensive process for lignocellulosic digestion coupled to biofuel conversion would have clear advantages over current pretreatments and the ethanol industry.

C acetobutylicum can consume lignocellulose (source), a wide variety of simple and complex carbohydrates, and can synthesize most amino acids from a suitable nitrogen source. It has long been valued as an industrial solvent producer and has a set of advance genetic tools for metabolic and biosystems engineering. Moreover, it is a native producer of butanol, an advanced biofuel. Butanol and its analogues hav been identified as prime targets for biofuels research, currently investigated by DuPont, BP, Gevo, and others. This microbe meets the requirements for low-cost non-food feedstocks and infrastructure compatibility. Therefore, C. acetobutylicum is an excellent chassis for an integrated biorefinery.

Clostridium acetobutylicum is a species historically used for solvent production in the 20th century.(REVIEWS) It belongs to a genus of anaerobic soil bacteria of industrial and pathogenic importance. It is known as the model for the Acetone-Butanol-Ethanol(ABE) fermentation, notably used in the Weizmann process of the first World War (REVIEWS). The Weizmann process produced 6 parts of butanol, 3 parts acetone, and 1 part ethanol in an anaerobic fermentation of starch, molasses, and other substrates. Its feedstock flexibility made it an excellent microbe for production during the first and second World Wars. Prized for its unique fermentation and tolerance of a variety of inputs, this organism remained the top producer of short-chain alcohols and solvents for decades, until the development of improved petrochemical processes.

Current research efforts are reviving this fermentation process with new genomic tools and analyses. The rate of genome sequencing efforts (NCBI stats) has resulted in (32?) sequenced Clostridia, providing insight into the metabolism and systems of the interesting species in this genus. Modern high-throughput sequencing has provided information about C. acetobutylicum's relatives and genetic tools (review) have enabled metabolic and biosystems engineering efforts in this genus. Metabolic engineering efforts have been made to increase the carbon flux towards the primary products.(Examples?) Efforts have been made to increase the carbon yield towards 1.(S.Y. Lee?) Attempts have been made to optimize the lignocellulosic machinery (source?) or to introduce alternative metabolic machinery (glycerol, wood ljungdahl).

Some attention has also been devoted towards increasing productivity of ABE fermentations by phenotype engineering. Acetone, butanol, ethanol, and other microbial waste products are produced during fermentation. These hydrophobic and amphipathic metabolic products intercalate cellular membranes and facilitate protein misfolding. This disrupts the electrochemical gradients of these membranes, leading to increased energy demands and changes to membrane stability. Also, increased protein misfolding affects nearly every process of the cell. Accumulation of these products induces a stress response to counteract the harmful effects of these metabolites. (sources) The systems responsible for solvent tolerance are complex and many specific responses remain poorly understood or unknown. The understanding of these systems is the goal of this study and I will briefly review the stress response in bacteria, focusing on its mechanisms and responses.

2.2 Bacterial Stress Response and Solvent Tolerance

Bacteria respond to a wide variety of environmental challenges through stress response systems. These networks of genes respond to signals indicating a hostile or desolate environment. Both gram negative and gram positive bacteria have developed physiological and genetic adaptations to these stressors. These adaptations generally slow growth rate, conserve resources, and prolong the viable life of the cell. Conditions of plenty in the natural environment are rare (source), and cells have developed adaptations to specific stimuli of the natural environment. Thus these adaptations are capable of detecting specific stressful conditions and transforming the signal into the activation of stress response systems, network of genes that enable cells to adapt and survive. Many stressful conditions exist, yet they all induce both specific and general stress responses.

Specific stress responses begin with the detection of a stressful environment. The absence of certain amino acids, minerals, vitamins, or energy may impair the cells ability to grow or survive. Cells can detect energy levels through intracellular messengers (and voltage gated ion channels?), activating programs that conserve resources and slow growth. Dangerous environmental conditions such as temperature fluctuations, oxidative and osmotic stress trigger other response systems that protect the cells from protein, DNA, or chemiosmotic gradient damage. Finally, cells also encounter biological, inorganic, and organic toxins that activate stress response programs to encourage proper protein folding, toxin export, and more.

There is one thing that most of these programs have in common: the general stress response. Many of these specific response programs also induce a general stress response. These programs provide non-specific survival benefits to others stressors and constitute many protective and conservational responses to environmental stress. Common objectives of stress response are met through the activation of general stress response programs. For example, during both nutrient deprivation and acid stress, cells must slow or cease growth to cope with changing energetic demands and the metabolic burden of the activation of both the specific and general stress response program. Proceeding with this example, I highlight one of the general stress response programs.

The stringent response may begin with a decline in the availability of an amino-acid or mineral (e.g. phosphorous, bioavailable nitrogen) in the environment. Decreased levels of the amino acid are observed in a cell, typically by riboregulators/riboswitches. These activate regulatory

Next, ppGpp accumulation modulates the expression of a large number of genes involved in primary metabolism, growth, replication, and competence (survival). This regulation is mediated in part by the RNA-pol holoenzyme itself and with some help from dksA (what is the role of dksA?). The stringent response was also shown to crosstalk with the Csr carbon-storage pathway to modulate the metabolic rate.

The general or non-specific stress response is induced after a specific stress response has begun. After detecting the stress, signal transduction events activate the general stress response system. Currently, the general stress response is divided into four classes of genes. The first class of genes is governed by the repressor HrcA. When the cell is stressed by one or more types of compatible stresses, the HrcA regulon is derepressed, increasing the expression of proteins such as the dnaK and groEL chaperonins. The second class of genes is

The general or non-specific stress response is induced by a variety of stimuli, such as energy starvation, or mineral, nutrient, or amino-acid deprivation. Specific signals activate two component signaling systems in the cell, which then relay the signal to response regulators or transcription factors. Specific and non-specific programs are thus activated, increasing the tolerance of the cell to its environment. An interesting finding (source) in the (1990s??) was that specific stresses produced non-specific survival benefits to other types of stresses. These non-specific tolerance systems constitute the general stress response. A typical example of this response include the so-called stringent response, which inhibits replication, reduces metabolic rate, and conserves energy.

A second system that is induced by multiple conditions and produces survival benefits to more than one type of stress is the heat-shock response.

The heat-shock response is most induced by conditions of high temperatures

(¿ 40 degrees) but may also be induced by osmotic, starvation, or other conditions (sources). This conserved system includes both specific and non-specific members (1990 general stress response of bsub). Canonical examples in this sample include the GroEL chaperonin machine (source) and proteases. The GroEL chaperonin system is a highly conserved group of proteins, forming a multimer. Through cycles of binding, ATP hydrolysis, and structural changes, the chaperonin proteins guide malformed proteins to their unfolded state and back again to encourage proper folding. Protein misfolding is increased not only under heat stress, but in solvent and pH stress as well. The heat-shock system is therefore particularly useful for biotechnological applications.

GENERAL STRESS RESPONSE A recent experiment showed an increase in final fermentation yield (productivity) by overexpressing certain stress response proteins and chaperones. These efforts follow engineering efforts in other organisms (e.coli source??). The increased productivity of these strains suggest that an abundance of chaperonins, proteases, and other stress response proteins may counteract the incrased protein misfolding rate due to abundant solvents and acids.

SPECIFIC STRESS RESPONSE

2.3 History of Genomic and Transcriptomic Research

Chapter 3
METHODS

3.1 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[1]. 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 OD 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.

3.2 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 with each addition. 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 transferring 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. The samples were moved to fresh 1.5 µL tubes.

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.

3.3 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 (1UµL^{-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 76 cycles over two lanes of an Illumina HiSeq 2500 at the University of Delaware Sequencing and Genotyping Center (Newark, DE, USA).

3.4 Sequencing Statistics, Alignments

Paired-end sequencing resulted in 749,709,771 pairs of 76 bp 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, the fastq headers are briefly pre-processed for downstream applications. Then, remaining sequencing adaptors were removed from the reads with Trimmomatic[2]. Base quality is adjusted by trimming to the minimum base quality of 20. Before aligning to the *Clostridium acetobutylicum* ATCC 824 genome, the data is subjected to *in silico* ribosomal RNA removal by aligning the reads to the rRNA sequences with Bowtie 2.1.0[3]. The unmapped reads were then aligned to the genome and megaplasmid sequences (NC˙003030.1 and NC˙001988.2). The alignment files were then cleaned, sorted, indexed, and validated with SAMtools[4] and Picard[5].

3.5 Coverage analysis

Coverage vectors for each strand were calculated with BEDtools[6] and summarized in R. Coverage vectors for each transcript were then acquired with a custom Ruby script. Summarization and visualization of these data was performed in R[7], circus, and d3.

3.6 Transcriptome Assembly and Annotation

Reference and *de novo* assembly was done with Trinity[8]. Fastq files were modified by appending the second column of the fastq Casava 1.8+ header to the first column before processing and alignment. Next, the resulting alignment files were merged and sorted before appending the pair information (/1 or /2) according to the Trinity documentation. To assess the assemblies, I have contributed to a transcriptome assembly assessment software project: Transrate. This software assesses transcriptome assemblies by calculating general assembly statistics, coverage statistics, and agreement with the reference proteome. I have made several additions to this software. Specifically, unpaired reads and strand specific alignment were integrated into the coverage/alignment statistics. Additionally, singleton reads were produced from the alignment process and were then further assessed for possible sources of contamination. Finally, the assembly itself was aligned to the reference genome, assuring the validity of the assembly and the identity of the assembled transcripts. The assembly, in fasta format, was then aligned to the genome with blat, converted to bed format, processed, converted to genePred and ultimately to gtf format. The gtf format assembly was then combined with the reference proteome for comparison. The transcriptome assembly was annotated using reciprocal best blast, TMHMM, HMMer, SignalP, TransDecoder, and otherS???? The reference proteome was compared with the assembly for the number of proteins, the OHR, and the differences between the annotations. Next, a statistical analysis of transcriptional start sites was performed with TSSi[9] after extracting (200? 300?) bp windows surrounding the beginning of each assembled transcript with a custom Ruby script. Coverage vectors from each transcript were visualized with these results in R.

3.7 Digital Gene Expression, Principal Components Analysis, and Differential Expression

Read counts per transcript were quantified with HTSeq[10]. Raw count data were visualized and normalized in R. The data were regularized following the conservative approach of DESeq2[11],[12]. The processed data was subject to Principal Components Analysis using the rgl library in R, and results were added to an interactive webpage. A Wald test was used to test for differential expression. Calculations and visualizations were done in R with various packages(OTHERS)[13]. Data were also processed manually for visualization in Circos graphs[14].

3.8 Gene Expression Clustering and Visualization

Regularized data were normalized or converted into Kendall, Pearson, and Spearman correlation matrices in R. The data were used as input to a parameter sweep with my hrefhttps://github.com/MatthewRalston/OPTICS-Automatic-Clusteringimplementation of the OPTICS clustering algorithm. I have tweaked the source code of the automatic feature extraction to be closer to the original algorithm (REFERENCE). Additionally, I added per-cluster metrics, described in the project README. This allowed me to visualize the results of the parameter sweeps and select the data input and parameter values that produce the best clusters. Exploratory data analysis was done with 'knitr' for HTML report generation.

3.9 Web Content

Interactive web material was generated using a mixture of javascript and HTML. The d3 library was used for dynamic content updating and interactivity. Circos was used to generate the larger circular plot visualization. These web pages are hosted on github for access by collaborators.

3.10 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 4

RESULTS

Chapter 5

SEQUENCING

5.1 Transcriptome Assembly

5.1.1 Assembly Quality

5.1.1.1 Quality Statistics

5.1.1.2 Example Transcripts

To assess the data/assembly quality further, I used canonical transcripts of *C. acetobutylicum* for comparison. Six issues, listed below, were considered for each example to better understand the quality of the assembly and the degree of curation required.

1. Is the transcript large enough to include the known ORFs and RBSes?
2. Does the assembled transcript's TSS agree with promoter motifs?
3. Does it agree with published transcription start sites?
4. Does the assembled transcript's size agree with published Northern blots?
5. Does the assembly represent the coverage and if not, which best represents the biological knowledge of this region?
6. Does the assembled region require curation (e.g. fused or truncated transcripts)?

Agreement between the data and the literature would support the efficacy of this technique. Furthermore, these results could be widely applicable to future studies if only minimal curation is required. The first example that I examine is the Sol locus.

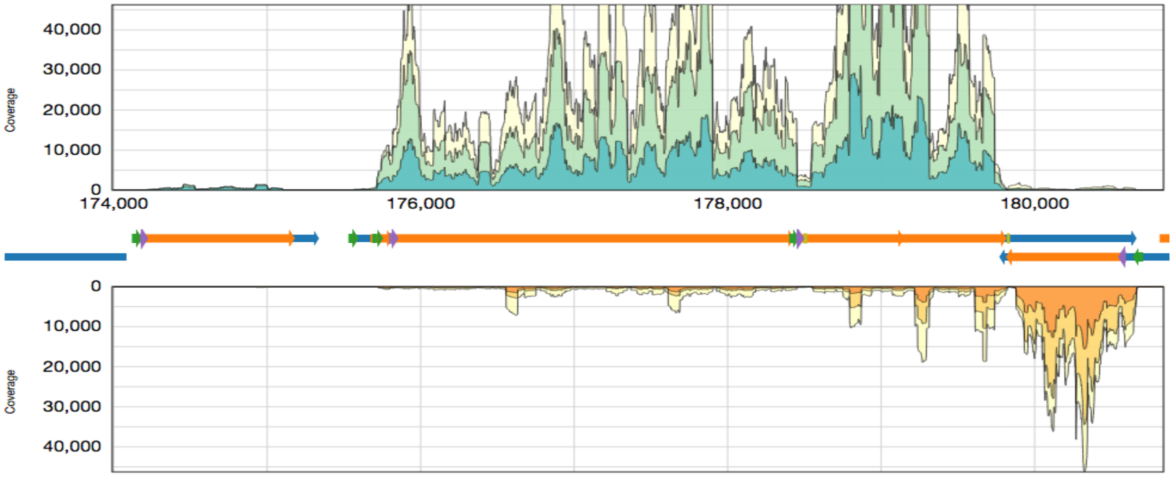
The Sol locus is a 5.1kb region(175530-180,650) on the pSol1 megaplasmid that is responsible for the production of several solvents. This region encodes several enzymes including a tri-functional NAD(H^+)-dependent alcohol/aldehyde dehydrogenase, two coenzyme-A transferases, and a acetoacetate decarboxylase. These genes are vital for acid reuptake and conversion into alcohols, a vital part of this organism's metabolism. In this dataset, we observe strong coverage (¿ 10,000x) of the Sol operon and ADC. The Sol operon has continuous coverage across the entire 4.1kb transcript and its size agrees with the literature. The depth of coverage is also fairly consistent, except for a 100bp region near the N-terminus of CtfA. This decrease in coverage can be explained by an inverted repeat present in this region, which could be difficult to sequence. It is also interesting to note that a promoter motif TTCATA(13)TATAAT is also present in the region, upstream of the RBS. It is interesting to note that all primer extension studies for this Operon used probes from AdhE for obvious reasons. To my knowledge, most Northern blot protocols follow the work of Durre et al, where probes were only designed for AdhE. In one source, a faint band is visible for a 2.6kb RNA, which matches reasonably to the 2.7kb continuous sequence that we observe in this example. Unfortunately, many other studies of this locus use Adh probes exclusively, primer extension, or have semi-quantified their blots, preventing further investigation of alternate bands. The assembly suggests a distal transcription start site at 175,564, agreeing with two previous studies of this region. Several increases in coverage are observed just after the proximal promoter as well. It is widely accepted that the proximal promoter is the most active promoter motif for the Sol operon and this view is supported by our data. The assembly also shows a distinct transcription start site for SolR, in agreement with previous findings. The coverage data also agree with the single transcription start site for Adc, shown as a distinct increase in this dataset. However, the Adc transcript has been fused to residual signal upstream of the true TSS. This is the first example of misassembly that I address in the next section.

The next example

5.1.2 Identify and Attempt to Resolve Remaining Issues

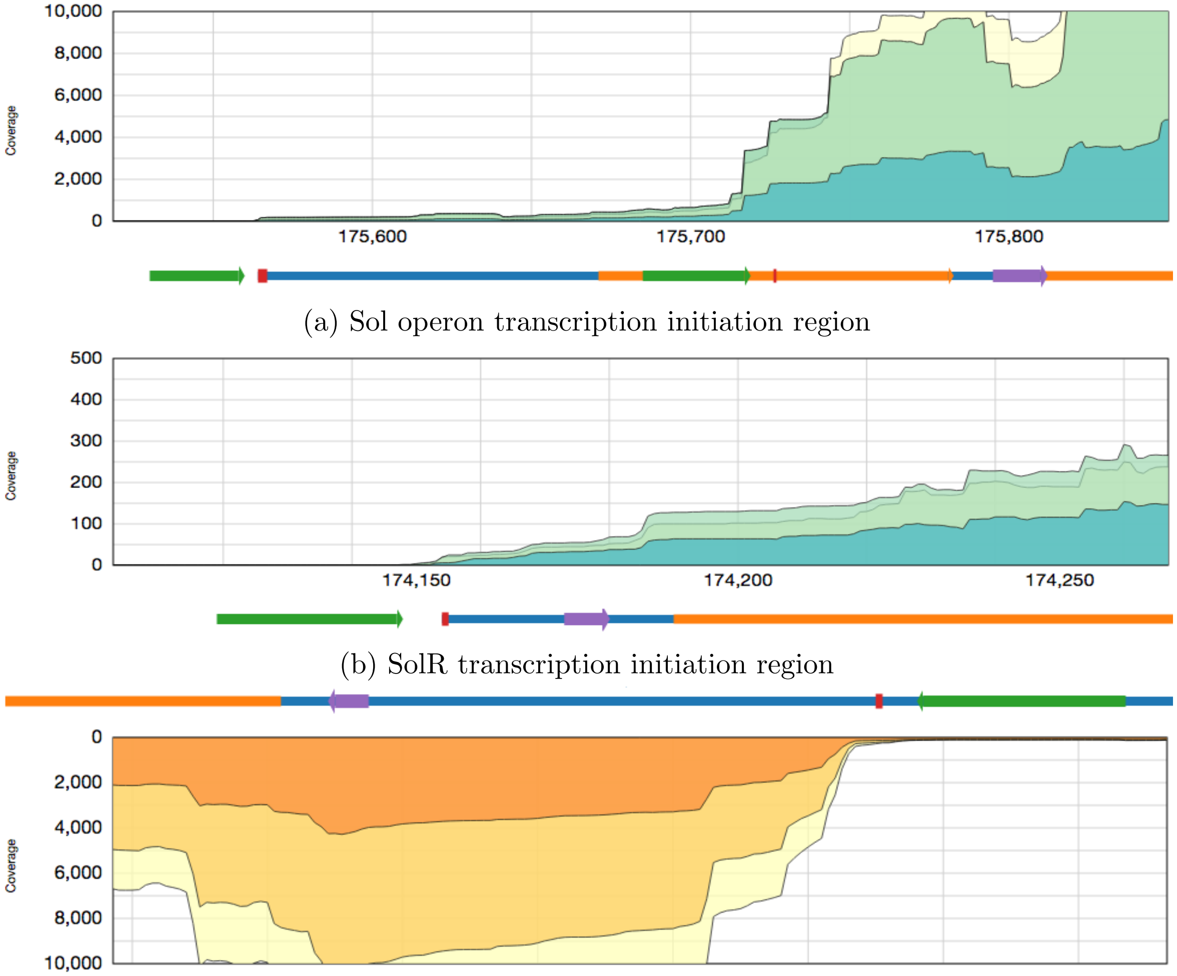
5.1.3 Novel Transcripts

5.1.4 Exploratory Tools



(a) Sol locus

Figure 5.1: Sol Locus: The Sol operon (a) upper track) consists of OrfL, alcohol dehydrogenase (AdhE), and Co-A transferases A and B (ctfA,ctfB). Acetoacetate decarboxylase (ADC; a lower track, right) is also shown. Coverage for the Watson and Crick strands (top and bottom tracks) are visualized with an annotation track (center). Tracks show cumulative coverage for unstressed (yellow), butanol (light green/ light orange), and butyrate (green/orange) stressed samples over all time points. Transcripts (blue), ORFs (orange), RBSes (purple), inverted repeats (yellow), promoters (green), and TSSes (red) are represented as arrows and bars.

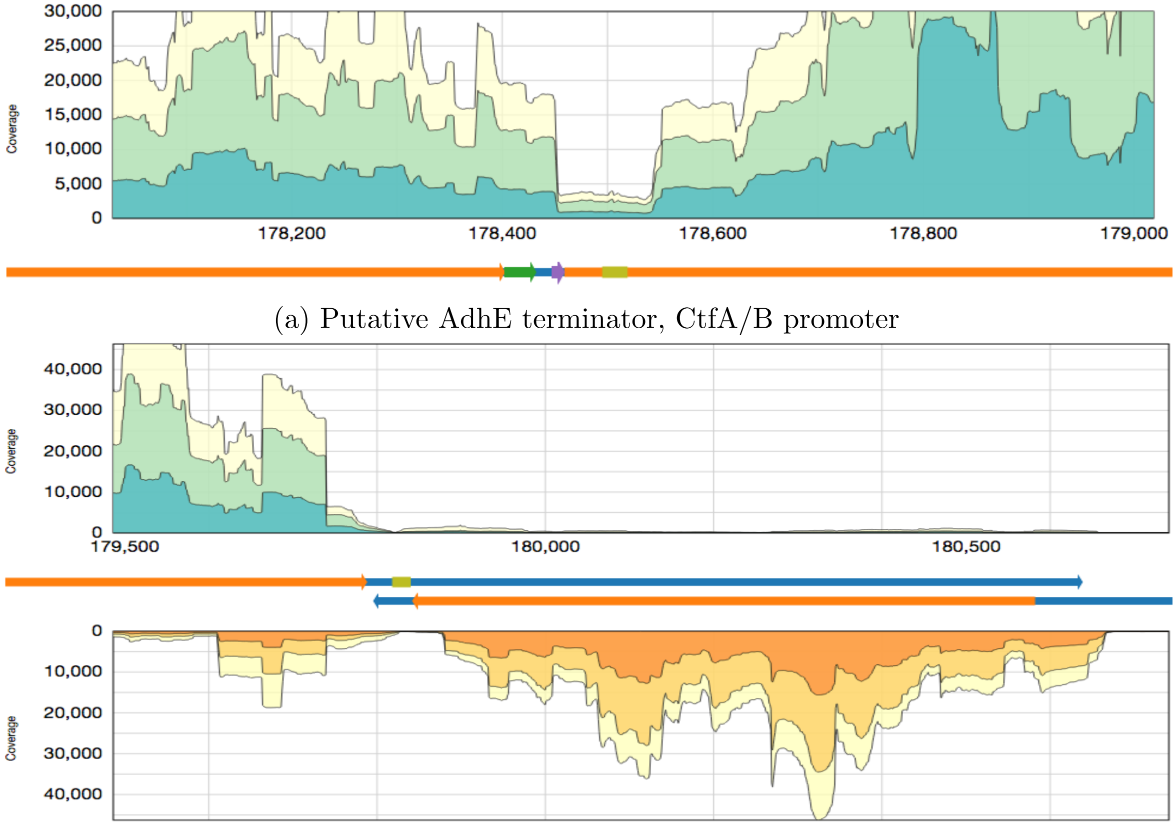


(a) Sol operon transcription initiation region

(b) SolR transcription initiation region

(c) Adc transcription initiation region

Figure 5.2: Sol Locus Transcription Start Sites: . a) Sol operon (OrfL, AdhE) transcription start sites. The coverage and assembly data have strong agreement with previously described proximal and distal promoters and transcription start sites. b) The assembled transcription start site for SolR agrees with previous findings. b) Transcription initiation region for Adc. While the coverage clearly shows the appropriate increase, the transcription start site has been fused to residual coverage upstream of the true TSS.



(a) Putative AdhE terminator, CtfA/B promoter

(b) Bifunctional Rho-independent terminator for Sol operon, Adc

Figure 5.3: Sol Locus Transcription Stop Sites: a) Low coverage in the Sol operon. A terminator may be partially responsible for a sustained low coverage level in the Sol operon. Additionally, a promoter motif was located upstream of the CtfA RBS and the pattern of expression is consistent with the beginning of a new transcript. b) A bifunctional terminator, likely responsible for the decrease in coverage observed for both transcripts on their respsective strands

Chapter 6

DIFFERENTIAL EXPRESSION

Chapter 7

DISCUSSION

Chapter 8

CONCLUSIONS

Bibliography

[1] 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.

[2] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page 170, 2014.

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

[4] 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.

[5] Alec W. Picard, 2009.

[6] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[7] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[8] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[9] 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.

[10] Simon Anders. Htseq: Analysing high-throughput sequencing data with python. *URL http://www-huber. embl. de/users/anders/HTSeq/doc/overview. html*, 2010.

[11] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *bioRxiv*, 2014.

[12] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[13] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2009.

[14] et al. Krzywinski, Martin. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

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

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

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

[24] 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.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

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

[35] 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.

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

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

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

[53] 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.

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

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

[56] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, page bbs017, 2012.

[57] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[58] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in rna-sequencing data. *Bmc Bioinformatics*, 12(1):290, 2011.

[59] Shawn ONeil and Scott Emrich. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC genomics*, 14(1):465, 2013.

[60] Grainne Kerr, Heather J Ruskin, Martin Crane, and Padraig Doolan. Techniques for clustering gene expression data. *Computers in biology and medicine*, 38(3):283–293, 2008.

[61] Barack Obama. Remarks by the president at u.n. climate change summit, 9 2014. Remarks by the President at U.N. Climate Change Summit at the United Nations Headquarters, New York, NY. [Accessed: 2014 09 30].