ASSEMBLING IMPROVED GENE ANNOTATIONS IN *CLOSTRIDIUM ACETOBUTYLICUM* WITH RNA SEQUENCING

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

Winter 2015

ASSEMBLING IMPROVED GENE ANNOTATIONS IN *CLOSTRIDIUM ACETOBUTYLICUM* WITH RNA SEQUENCING

by

Matthew T. Ralston

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

Approved: _____
Errol L. Lloyd, Ph.D.
Chair of the Department of Computer Science

Approved: _____
Babatunde A. 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 has been critically inspirational for this work's focus on sustainability and climate change. I cannot emphasize how much she is the foundation of my views on the purpose of research and development. With special thanks to my Father, for inspiration through his work ethic, leadership, and open mind. In difficult times, he has always had faith and patience for me and I would not be in this program without him. With special thanks to my sister for her support of my education, curiosity, and maturity. I am eternally grateful for her sense of justice and duty; I would not be here without her support, 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 unconditional love. I write this with special thanks for my grandfather George for inspiring my pursuit of science, chemistry, and music. With special thanks for my grandmother Winnie for the inspiration of her love and the optimism which bring me through each challenge. With special thanks for my uncle Kevin for his inquisitive nature, friendship, optimism, and support. He has found creative ways to teach me about science and pioneering spirit even as busy as he is. With special thanks for my brother-in-law Coy for the friendship, encouragement,

and inspiration he has shared with me; I greatly admire his sense of honesty, selflessness, and humor that have made him an *excellent* role model for me. With special thanks for my nieces Violet and Ruby for their laughter, love, and their infectious energy, ingenuity, and optimism. Many thanks for girlfriend Madeline for her love, open ear, curiosity, support, and encouragement throughout the years. She has an unshakeably good heart and a responsibility towards others, which truly *inspire* my belief in how my work can touch the lives of others. Many thanks to my best friend Andrew for his friendship, curiosity, support, and encouragement. A man with character and grit, we share a unique bond that has always made us a good team and great friends. 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. This would not be possible without the kindness, consideration, and respect that Karol has shown me and 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. Her drive, honesty, attitude, and beaming personality encapsulates who I've always believed researchers can be. With thanks to Rachel Marine, a great and inquisitive scientist and friend who always has insight and words of encouragement. With thank to Ryan Moore for his openness, honesty, encouraging reasoning, sense of humor, and of course 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 faith throughout this project. The success that we've seen in this effort I owe to Keerthi's guidance and training. They are without a doubt the best scientific mentors I have ever had the pleasure of studying under. I look forward to every meeting that we have and it will be difficult to depart from their guidance when we conclude this project. 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. There is something unique about the Papoutsakis group. Each and every member is innovative, industrious, and full of excellent questions. Sharing lunch, riddles, cake, and time with this lab has been extraordinarily fulfilling. For the hard work, late nights, early autoclave cycles, shared spaces, and plenty of reasons to celebrate I am grateful for each and every member in this group. Because of this team of excellent scientists and engineers, this work was guided to success. Thank you, all.

Ad maiorem dei gloriam.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The *C. acetobutylicum* genome annotation has been markedly improved by integrating bioinformatic predictions with RNA sequencing(RNA-seq) data. Samples were acquired under butanol, butyrate, and unstressed treatments across various growth stages to sample the transcriptome from a range of physiologically relevant conditions. Analysis of an initial assembly revealed errors due to technical and biological background signals, challenges with few solutions. Hurdles for RNA-seq transcriptome mapping research include optimizing library complexity and sequencing depth, yet most studies in bacteria report low depth and ignore the effect of ribosomal RNA abundance and other sources on the effective sequencing depth.

In this work, workflows were established to address type I and II errors associated with these challenges. An integrative analysis method was developed to combine motif predictions, single-nucleotide resolution sequencing depth, and library complexity to resolve these errors during assembly curation. This contextualization minimized false positive error and determined gene boundaries, in some cases, to the exact basepair of prior studies. Curation of the pSOL1 megaplasmid reconciled transcriptome assembly statistics with findings from *E. coli*.

The resulting annotation can be readily explored and downloaded through a customized genome browser, enabling future genomic and transcriptomic research in this organism. This work demonstrates the first strand-specific transcriptome assembly in a *Clostridium* organism. This method can improve the precision of transcript boundary estimates in bacterial transcriptome mapping studies.

Chapter 1

INTRODUCTION

1.1 Motivation

Increases in global CO2 levels, sea level, temperature, and acidification are tipping climate models toward disaster.[1] Few solutions exist for what has been described as the "...issue that will define the contours of this country more dramatically than any other."[2] A chief issue with the global CO2 equation is the lack of systems that utilize this greenhouse gas. A renewable chemicals industry has been suggested to restore balance to our climate system in an economically sustainable way.

Leading scientists and engineers[3;4;5;6;7] recognize *Clostridum acetobutylicum* as a potential platform organism for a biorefinery, a bioprocess featuring an organism which converts different substrates into chemicals. This microbe produces an advanced biofuel called butanol, an energy-dense gasoline replacement. Recently, genetic tools have been developed to optimize *C. acetobutylicum* productivity.[8;9;10] Metabolic engineering techniques have been investigated[11;12] to increase butanol yield and its already impressive feedstock flexibility.[13;14] However, successful strains also require biosystems engineering to increase robustness and fuel tolerance.[13;15;16;17] Limited knowledge of the biofuel-stress tolerance systems in *C. acetobutylicum* is a barrier to the development of robust strains for renewable fuel production.

Research and development in *C. acetobutylicum* requires a complete and experimentally determined genome annotation. A complete set of transcript boundaries

would facilitate research into the biochemical and molecular systems responsible for biofuel tolerance. For example, transcription start sites (TSS) would enable the discovery of regulatory motifs and network structure responsible for the solvent-tolerant phenotype.[18;19;20;21] Only open reading frames were predicted in the original genome annotation that have not been verified by absolute expression measurements.[22] To provide the features of interest, a transcriptome mapping study was designed.

Transcriptome mapping studies frequently utilize high-depth Next-Generation Sequencing methods to provide sensitivity for low abundance transcripts and their features.[23;24;25;26] However, the required depth of sequencing for high-sensitivity bacterial transcriptome mapping is unknown[27;28;29]. Authors frequently report biased total read counts or fold-coverage estimates instead of per-base sequencing depth.[30;31;32] Moreover, studies often fail to quantify ribosomal RNA removal rates[30;31;32] which have a dramatic effect on effective sequencing depth.[33;34;35] Additionally, these studies rely on sequencing depth alone, ignoring the complexity of the dataset for assembly.[25;26;36] Without the empirical guidelines described for Eukaryotic applications,[27] these studies under[30;37] or over-estimate[31;32] the required amount of sequencing yet observe low depth and complexity[32] and poor utilization(Table 2.1) of their sequencing results.

1.2 Approach

To facilitate future renewables research in *C. acetobutylicum*, an RNA sequencing(RNA-seq) transcriptome mapping study was designed with sensitivity in mind. A mixture of laboratory and informatic approaches were used to identify and remove "noise" (e.g. rRNA, duplicate reads) from the sequencing library, leading to a high-quality

sequencing dataset and unbiased sequencing-depth statistics. Transcriptome assembly was used to describe operon structure and estimate transcript boundaries. Finally, a genome browser was constructed to identify and resolve misassemblies and share the data with the *Clostridia* research community.

1.3 Document Overview

This thesis describes a technique for unbiased and precise estimation of transcript boundaries in microbial systems. The document compares related approaches for transcriptome mapping and identifies their strengths and weaknesses. A methodology is described that lead to a high-quality sequencing dataset, optimized with considerations of biological and technical noise. After describing this technique, the sensitivity of the technique is qualified both in terms of fold-coverage and sequencing depth. Next, an assessment of the assembled dataset is presented, which reveals challenges associated with high-depth sequencing, either not detected or described by comparable studies. This assessment lead to the development of a genome browser that was used for proof-of-principle curation of the pSOL1 megaplasmid, markedly improving the false positive or type I error rate. This document describes a method and its application in the model solventogenic bacterium *C. acetobutylicum*, marking the first reported strand-specific transcriptome assembly in the *Clostridia*. This thesis describes the discovery of novel genomic and transcriptomic features that will be shared through the genome browser with the entire *Clostridia* and renewables research community.

Chapter 2

BACKGROUND

2.1 Renewable Chemicals: A 21st Century Challege

Climate change research[1] and petrochemical exploration[38;39] suggest that escalated weather variation, sea levels, and atmospheric and oceanic temperatures will accompany steep increases in the cost of petroleum. Renewable chemical platforms will be an increasingly economical solution to climate change. Renewable biofuels can be carbon neutral sources of energy. Renewable biochemical processes can actually behave as carbon sinks, with net accumulation of CO_2 in chemicals and biomass after subtracting processing energy requirements.

Many renewable biochemical systems revolve around a biocatalyst, a microbe that can convert low cost substrates into fuels or chemicals.[3;4;5;6;7;40] This production system, frequently referred to as a "biorefinery," requires a microorganism with a wide range of potential feedstocks and a natural biofuel producing metabolism. Biofuels with comparable energy density and hygroscopicity to conventional fuels are desirable for infrastructure compatibility. The butanol-producing bacteria *Clostridium acetobutylicum* consumes a number of sugars and hydrolysates[13] and produces butanol, a direct gasoline replacement.[41]

C. acetobutylicum is a historically industrial solvent producer.[42] It consumes hemicellulose, a variety of simple and complex carbohydrates, and hydrolysates.[13;42] Also known as the Weizmann organism, *C. acetobutylicum* converts these substrates

into solvents through an acetone, butanol, and ethanol (ABE) fermentation.[13;42] Importantly, these cells synthesize most amino acids with ammonium salts as a nitrogen source, requiring only a minimal defined medium.[43;44] This microbe meets the requirements for low-cost non-food feedstocks and infrastructure compatibility. Therefore, *C. acetobutylicum* is an excellent chassis organism for an integrated biorefinery and is the system of study in this work.

C. acetobutylicum has a number of intrinsic advantages that minimize the engineering efforts required for bioprocess development. It is one of over 17,000 bacteria with sequenced genomes.[22;45;46] For example, current metabolic models[12] are used for sophisticated metabolic analyses, such as C^{13} metabolic flux analysis.[11] A model for the solventogenic *Clostridia*, *C. acetobutylicum* is a reasonably well studied organism with industrial potential.

Prior to the genomic era, targeted studies in this organism revealed the specific loci for solvent formation,[47;48;49;50;51;52;53] sporulation,[54] and canonical heat-shock operons.[55;56] These genes were typically cloned, sequenced, and investigated with gene-specific transcriptomic techniques. However, only the mechanisms of the unique metabolic systems (e.g. solvent formation) have been investigated in detail; many proteomic and transcriptomic mechanisms behind the *C. acetobutylicum* regulatory networks remain unknown.[57;58] The mechanisms for the majority of the genetic and metabolic systems in *C. acetobutylicum* are largely inferred from homology, often an appropriate assumption.

That being said, many interesting characteristics of *C. acetobutylicum* are unique to solventogenic *Clostridia*. Of particular interest for renewable fuel research is its solvent stress-response, which may be uniquely adapted to its solvent-producing metabolism. A number of stress-response systems exist for specific stresses while broader systems can respond to multiple stressors. By exploring the knowledge

of these systems in *C. acetobutylicum*, gaps in understanding can be identified for this work to explore. In the next section, stress-response systems are reviewed, demonstrating opportunities for the discovery of novel transcriptomic features.

2.2 Biofuel/Solvent Tolerance and the Bacterial Stress Response

Bacteria respond to a wide variety of intrinsic and extrinsic challenges with stress response systems.[6;59;60] For example, nutrient deprivation, osmotic shock, temperature fluctuation, and high chemical concentrations are common in their natural environments.[59] Cells activate specific or general response networks after these insults. *C. acetobutylicum* is a naturally solventogenic bacterial species and could possess biofuel or solvent-tolerance genes that would likely be regulated by general or specific stress response systems. Such genes would be natural targets for biosystems engineering, improving the tolerance of engineered strains to high biofuel titers.

Unfortunately, knowledge of these systems is incomplete in *C. acetobutylicum* and no unique solvent tolerance mechanisms, such as solvent exporters, have been identified. The stress response is an important system for biotechnology, but the annotations of *Clostridia* genomes require improvement for understanding of these and other systems. Here we review these systems and their mechanisms to understand how a solvent stress response system would function.

2.2.1 Specific Stress Response Systems

Specific stress-responses are designed to mitigate the negative effects of a particular stressor. These systems typically contain a detection mechanism for a molecular stressor, such as antibiotic compounds,[61] or its effects, such as DNA damage.[62]

Antibiotic resistance is an example of a specific response system that detects a molecular stressor. In the presence of organic compounds of the β-lactam[61] or tetracycline families,[63] bacteria activate antibiotic resistance genes that either export or modify the organic compounds to prevent their action. Some of these antibiotic-resistance genes are part of operons that possess antibiotic-detecting repressors.[61] Upon detection of the antibiotic agent, a conformation change triggers derepression of the antibiotic resistance gene, resulting in an antibiotic resistant phenotype. This stress response system helps bacteria to quickly and specifically respond to a family of antibiotic compounds. Interestingly, multidrug resistance genes can contribute to solvent tolerance in *P. aeruginosa*[64] and *E. coli.203*

A second example of a specific stress response is from *D. radiodurans*, which has an unparalleled resistance to ionizing radiation.[65] In response to breaks in DNA, *D. radiodurans* repairs the multiple copies of its genome, enabling growth, viability, and survival at over 5,000 Gy of radiation.[65] This system responds instead to DNA damage, a symptom of γ-radiation. The extreme tolerance of *D. radiodurans* to radiation is a byproduct of a specialized system for DNA repair,[65] augmented from the standard DNA repair system.[62] This system contains a detection system, non-homologous recombination, and a specialized DNA repair system allows *D. radiodurans* cells to survive and repair hundreds of insults to its genome.

If specific stress response systems exist for solvent stress (e.g. solvent-exporting efflux pumps) in *C. acetobutylicum*, they would be desirable targets for biosystems engineering. Solvent tolerance pumps, such as the SrpABC efflux pump of *P. putida*, could enable a butanol-tolerant phenotype.[66] A solvent efflux system was not described in the original genome annotation,[22] although gene models have changed significantly in the previous 14 years. An updated, improved genome annotation could enable the identification of solvent efflux genes.

In addition, stress responsive small RNAs (sRNAs) have been described,[44] although their roles, regulation, and conservation remain unknown. This discovery suggests that there are previously unknown active regions of the *C. acetobutylicum* genome that require investigation. Naturally, differential expression experiments are desirable to understand the dynamics of sRNAs, their targets, and other novel transcripts. However, the statistical treatment of measurements for such experiments is complex and requires reliable estimates of transcript expression.[28;29;67] While these estimates would be impossible to acquire with the original bioinformatically-predicted genome annotation, an improved genome annotation could benefit these investigations in several ways. First, transcriptome mapping could identify novel stress-responsive transcripts that could be co-regulated with the previously discovered small RNAs.[44] Novel stress-responsive transcripts could encode ORFs homologous to exporters or efflux pumps. Precise transcript boundaries could improve gene-expression estimates by counting reads at the transcript level, as opposed to the ORF level. Therefore, an improved annotation could reveal novel components of the stress response and would support differential expression investigations of these systems.

In addition, an improved genome annotation complete with transcription start sites and regulatory regions would facilitate research on stress-responsive genes and associated regulatory motifs. Regulatory motifs could be discovered with computational methods[68;69] with a complete set of transcript boundaries.[19;20;21] Solvent responsive regulatory motifs would be useful for designing a semi-synthetic stress response.[70;71] If *C. acetobutylicum* possesses a specific solvent-response system, an improved genome annotation could also enable the discovery of corresponding regulatory motifs.

The specific stress-response systems have unique detection and response mechanisms to particular intrinsic or extrinsic stressors. These genes have important applications in environmental remediation,[62] pharmaceutical research,[61] and renewables research. While a solvent stress response system has not yet been identified in solventogenic *Clostridia*, a genome annotation could reveal stress-responsive genes and transcripts, beyond those identified by ORF predictions alone. In addition to these benefits, an improved genome annotation could also benefit knowledge of general stress response systems and will be discussed next.

2.2.2 General Stress Response Systems

In contrast to specific stress-response systems, general responses are activated by more than one stimulus. For example, during both nutrient deprivation and acid stress, cells must slow or cease growth (stringent response) to adapt to energetic demands of the activation of both the specific and general stress response program. Also, both heat-shock and solvent stress can denature proteins and consequently activate chaperonin systems, another example of a general stress response.[60;72;73] After detection of the stressor, signal transduction events activate dormant response machinery or activate/derepress response systems[60;74]. These systems can be useful for stress response engineering[70;71] and are somewhat conserved across genera, although their knowledge in *C. acetobutylicum* remains incomplete. The general stress response is divided into four classes of genes based on the regulator responsible for their activation.

2.2.2.1 Class I

The first class of general stress response genes is governed by the repressor HrcA, which responds to protein denaturation from thermal or chemical causes.

The hrcA regulon contains at least 3 transcripts including the hrcA and dnaK/J, groES/EL, and htpG loci.[58] Denatured proteins titrate the GroEL chaperone from HrcA/GroEL complexes, resulting in a conformational change of HrcA and decreased DNA binding.[60;74] Operons regulated by the HrcA repressor are subsequently derepressed, rapidly increasing the amount of heat shock proteins. Protein denaturation negatively affects nearly every program and structure of the cell, resulting in decreased survival and viability. In *C. acetobutylicum*, the HrcA motif was recently described for standard heat-shock operons.[58] Additionally, class I genes are solvent-stress responsive due to solvent-induced protein denaturation.[72;73] It is unknown if any additional operons are also regulated by this repressor in *C. acetobutylicum*. To answer this question, an improved genome annotation including transcription start sites would facilitate the discovery of additional genes in the HrcA regulon through *in silico* analyses.

2.2.2.2 Class II

The second class of genes is regulated by a stress-responsive σ-factor, σ_B. In *B. subtilis*, the σ_B regulon consists coordinates a general stress response for a variety of stressors. A *C. acetobutylicum* σ_B ortholog was not predicted in the initial genome annotation,[22;58] although orthologous genes from its regulon have not similarly disappeared.[75;76;77;78] The regulation of these genes is unknown in this organism. Perhaps σ_B genes in *C. acetobutylicum* are under the control of overlapping regulons, as in *L. monocytogenes*[79] or are regulated by an unknown mechanism. Given the large size of the σ_B regulon and the presence of several of its genes (e.g. clpC[75]), the promoter and regulatory regions of *C. acetobutylicum* orthologs of σ_B-regulon genes would be useful for understanding the unique stress response of *C. acetobutylicum*. This class

of general stress-response genes present another opportunity for an improved genome annotation to facilitate stress response research.

2.2.2.3 Class III

The third group of genes are governed by the repressor CtsR, a dimeric helix-turn-helix regulator capable of responding to heat-shock, oxidative stress, and acid stress.[80] In *B. subtilis*, CtsR responds to heat-shock when ClpC, the fourth gene of the CtsR operon, releases McsB.[81] Free McsB phosphorylates CtsR, leading to positive autoregulation and derpression of the ctsR regulon.[81;82;83] This mechanism is thought to vary across the gram positive bacteria,[82] but the operon organization suggests that the *C. acetobutylicum* CtsR system is similar to *B. subtilis*.[58] A CtsR motif was found ahead of canonical CtsR regulon genes in *C. acetobutylicum*(Qinghua paper) although additional genes may be controlled by CtsR in the absence of σ_B. A genome-wide motif search in this organism could similarly reveal CtsR regulation of solvent responsive transcripts. Similar to the class I system, knowledge of the ctsR regulon would also benefit from an improved genome annotation.

2.2.2.4 Class IV

The fourth and final class of general stress-response genes are regulated by unknown mechanisms.[59] In *C. acetobutylicuM*, this class includes stress-responsive genes with unconfirmed operon structure and no confirmed motifs. Microarray experiments from Venkataramanan *et al.* show over 1,000 solvent responsive genes in *C. acetobutylicum*, the regulation of which remains almost completely unknown.[58] As suggested above, an improved genome annotation would aid the categorization of this massive gene set into class I, III, or potentially new regulons.

These general stress response programs are vital for adaptation and robustness. However, both sporulation and stress-response systems in the *Clostridia* differ from the model for sporulating gram positive bacteria, *B. subtilis*.[57] At the very least, key genes are missing such as the sporulation kinases and the stress response regulator σ_B.[57] Furthermore, most solvent responsive genes in *C. acetobutylicum* are regulated by unknown motifs and response regulators.[58] It is reasonable to expect that there may be some stress-response genes specific to the solventogenic *Clostridia*. Such genes would have had little homology to known genes during the initial genome annotation and therefore would not be included in previous comparative genomic and microarray analyses.[22;58] To illustrate the plausibility of this hypothesis, a recent study identified solvent responsive small RNAs in *C. acetobutylicum*, many with no known homologs or regulators.[44] Clearly, there is much to be done to understand and develop this organism for biofuel applications.

In this section, the example of stress-response systems was used to demonstrate differences of *Clostridia* from the *Bacillus* model. It is clear from this review of *C. acetobutylicum* stress response systems, there are opportunities to discover novel transcripts or proteins not described by the initial genome annotation and provide precise transcript boundaries for motif identification. To provide definition to the stress response and other systems, regulatory regions and operon organization should be revealed by modern transcriptomic techniques. Techniques that provide these details throughout the genome are discussed in the following section.

2.3 Transcriptomic Research

Modern high-throughput techniques can produce gloabl datasets, even when reference genomes are not available.[84] A large number of array and sequencing techniques have been developed to investigate the proteome, transcriptome, and

interactome including the elements of protein-protein,[85] protein-DNA,[86;87;88] protein-RNA,[89] and RNA-RNA interactions,[90;91;92] protein post-translational modifications,[92] single nucleotide polymorphisms,[93] and transcript expression.[94;95;96;97] These powerful genome-wide techniques allow experimental investigation of nearly every aspect of biological systems, including the identification of all genes and their boundaries. Here, we focus on transcriptomic techniques to identify these features.

Transcriptomic techniques allow the characterization of the properties and dynamics of RNA species. Microarrays and sequencing techniques have revolutionized transcriptomic research, providing new insights into the complexity of the transcriptome[95;97;98;99] and its regulation.[100] In addition to the measurement of specific populations[101;102], specific features of the transcriptome can be quantified.[103] Two common experimental approaches are used to explore cellular programs with high-throughput transcriptomic techniques: annotation and differential expression.

Before differential expression experiments are conducted, or in instances where a complete genome is not available, the catalog of all genes and their transcripts is required for microarray probe design or sequence read counting. ORF prediction algorithms[104;105] and annotation suites[106;107] can identify putative ORFs encoding enzymes and canonical genes for sequenced genomes, with small false positive and negative error rates. However, unique genes and small RNAs cannot be identified with these predictions. It is desirable to experimentally determine the transcriptome for conditions of interest.

The most common techniques for cataloging expressed transcripts and their properties are the tiling microarray[108] and deep RNA sequencing[32]. Microarray based analyses use probes spanning an entire genome, with some amount of overlap, to detect transcriptional activity.[109] Deep RNA sequencing[32] measures transcription

through the number of cDNAs sequenced from fragmented RNA, roughly proportional to the expression level.[67] Strand specific options for these methods are especially useful in dense bacterial genomes.[32;108] The large amount of data require computational processing to identify the desired features.[110;111] However, false positive and false negative errors (type I and II errors, respectively) from these techniques make automation of transcript annotation difficult, an issue infrequently discussed in the literature. Nevertheless, transcriptome assembly methods permit the reconstruction of full-length transcripts from sufficiently deep sequencing datasets at the expense of misassembly errors.[25] In this section, transcriptomic techniques are reviewed and experimental designs discussed to identify opportunities for improvement over convention for this study's objective, revealing transcript boundaries, regulatory regions, and operon organization.

2.3.1 Analytical Techniques

2.3.1.1 Microarrray

Microarray-based transcriptomics allows the simultaneous measurement of thousands of sequences simultaneously.[94] Probes designed for genomic sequences or open reading frames(ORFs) measure the expression of segments of the genome or ORFs, respectively. The former, tiling microarrays, are used to identify features of the transcriptome and annotate the genome. The latter is a typical practice for expression profiling experiments and identifying relative expression differences.[94;108] The experienced research community have made the microarray an excellent platform for transcriptomic analyses.

Microarray technology has limitations that should be considered during the early stages of experimental design. First, the microarray suffers from limited detection range and sensitivity compared to RNA sequencing.[112] Additionally, probe

design is limited to either ORFs only or to overlapping segments of a genome sequence where available.[96] Finally, the cost of tiling array experiments is a function of the amount of overlap between the probes; higher resolution implies higher cost.[112] The limited detection range could make low-abundance transcripts challenging to identify or distinguish from spurious transcription.[113;114]

A comprehensive investigation of *B. subtilis* identified many ORFs to be expressed under various environmental and life-cycle conditions.[108] The authors detected 85% of all previously known transcription start sites.[108] With 22-basepair resolution, tiling microarrays were used to detect increases in fluorescent signal across the genome, providing transcript boundaries after manual annotation. The use of manual methods to identify novel genes and transcript boundaries highlights the complexity of these datasets and the lack of methods for automated annotation. This study[108] is an excellent model for transcriptome mapping using the microarray technology.

2.3.1.2 RNA Sequencing

The emerging technology of RNA sequencing(RNA-seq) is displacing tiling arrays for transcriptome mapping and annotation studies.[112] Parallel sequencing platforms such as 454[115] or Illumina[116] enable the sequencing of gigabases of cDNA with precision for mapping and counting. cDNA libraries are sequenced in an extremely parallel manner, producing millions of sequenced "reads", which are then aligned to the genome. RNA-sequencing has a higher dynamic range than microarray technology in addition to basepair-level resolution.[112] These characteristics make RNA-seq optimal for precise determination of transcript boundaries, even for low abundance transcripts.

To achieve this, each step of the experiment and RNA processing must be considered carefully. Poor consideration of the factors involved with this method leads to uneven coverage,[25;26;27;28;32] poor depth,[31;32] and questionable strand specificity.[30] While attempts have been made to establish standards for biomedical RNA-seq,[27] no guidelines exist for the broader research community. RNA-seq is frequently performed with the Illumina platform for the amount of data it produces. This amount of data plays a crucial role in detecting transcript boundaries and low abundance transcripts.

A central challenge to the design of RNA-sequencing experiments is related to detection limit. The likelihood of sampling low-abundance cDNA fragments from transcript termini or low abundance transcripts is a function of the number of sequenced reads.[25;26;27;29] The absence of reads at a location in the genome indicates either an insufficient sequencing depth(false negative or type II error) or a true absence of transcription. Bacterial cDNA libraries are commonly sequenced in multiplex over one or more lanes of an Illumina sequencer, leading to millions of reads distributed across the libraries. Therefore, there is a three-part tradeoff between the replication, depth, and independent variables in the experimental design. A recent study concerning the tradeoff between the first two concluded that for differential expression experiments, replication is preferable.[29] In the case of transcriptome mapping and genome annotation however, additional depth may be preferable.[27]

Standard procedures are required to compare sequencing depth between RNA sequencing studies with similar objectives (e.g. transcription start sites). The Encyclopedia of DNA Elements(ENCODE) research consortium frequently uses RNA-seq in various forms[117;118] and has published best practice guidelines for Human genome research.[27] While average per-base sequencing depth and their distributions are not provided, they suggest that 100-200 million clusters of paired-end reads is sufficient

	Genome(Mbp)	Transcriptome(Mbp)	Clusters(M)	rRNA-free(M)	Mapped(M)	Alignment Rate	*Ratio*†
Standard[27]	3000	140	30	30	25	N/A	0.18
Deep[27]	3000	140	100	100	100	N/A	0.71
Ultra-deep[27]	3000	140	200	200	200	N/A	1.43
C. beijerinckii[31]	6	6-12	14	N/A	11.5	0.82	N/A
P. difficile[119]	4.3	4-8	50	4	N/A	N/A	0.5-1
B. anthracis[120]	5.5	5-11	33	N/A	5	0.15	N/A
H. pylori[30]	1.7	1.5-3	0.4	N/A	0.2	0.54	N/A
Synechocystis. sp[37]	3.9	3.9-7.8	0.2	0.1	0.1	0.475	0.03
E. coli[32]	4.6	4.5-9	52	N/A	17.7	0.34	N/A
P. gingivalis[121]	2.3	2.3-4.5	15	N/A	2.3	0.15	N/A
S. typhi[122]	5	5-10	5.7	N/A	1.8	0.31	N/A

Table 2.1: Study Comparison: Poor Data Utilization Rates
RNA-seq transcriptome mapping studies have widely different sequencing depths and alignment rates. These depths are not directly comparable between organisms; rather, the number of clusters/reads divided by the size of the transcriptome† is a more ideal metric for comparison. However, there is little discussion of the diluting effect of rRNA on useful sequencing depth, such as the 8% utilization in a study in *P. difficile*.[32] As a result, the ratio† cannot be calculated precisely for all studies. That being said, the poor read alignment and rRNA removal rates suggest that many studies do not achieve desirable sequencing depth given the size of the transcriptome.

depth to identify novel transcripts and transcriptomic features (e.g. TSSes) in the hg19 *H. sapiens* transcriptome. This number is much larger than required for microbial genomes; a preferable metric is the ratio of the number of reliably-mapped non-ribosomal reads to the approximate size of the transcriptome. Next, this metric is used for persepective to compare transcriptome mapping studies.

2.3.2 Transcriptome Mapping Studies and Common Challenges

The most phylogenetically similar organism to *C. acetobutylicum* that has been investigated with RNA-seq transcriptome mapping is *C. beijerinckii*.[31] In this study, 82% of the reads aligned uniquely to the genome, a good alignment rate and depth when compared to other studies (Table 2.1). However, the authors fail to qualify their work with respect to several factors that influence transcript discovery and transcription start site identification. These common issues to transcriptome

mapping studies are described next.

In dense bacterial genomes, proteins are coded by polycistronic transcripts, operons, that are packed closely in to small circular genomes, only a few megabases in length. In bacteria, overlapping transcripts can be encoded on opposite strands. Strand specific techniques are seldom used,[32] preventing the detection of divergent operons, novel genes in antisense, and *cis*-encoded sRNAs.

Unlike eukaryotic transcriptome mapping studies where poly-A selection is available, bacterial total RNA extracts contains 95-99% ribosomal RNA.[33;34;35] For most RNA-seq applications, the overwhelming presence of rRNA lowers the useful sequencing depth extraordinarily. Several commercial kits are available with suboptimal and inconsistent removal rates.[32;33;34] Very few studies provided calculations or discussion of the effect that rRNA had on the effective sequencing depth.

Another intrinsic artifact to the RNA sequencing method is the effect of preferential PCR amplification on library complexity.[123;124] After cDNA library construction, the library is typically amplified to provide additional material for sequencing. Duplicated reads provide redundant information and should be removed *in silico*.[125] It seems that no efforts were made to address this issue in these studies.

Ribosomal RNA and PCR-amplification bias are two sources of noise that lower the amount of useful signal from bacterial RNA-seq experiments. *In silico* solutions[125;126] allow researchers to separate and quantify the useful signal for the purposes of qualification and comparison with other studies. Ribosomal RNA was treated in two of eight studies reviewed, showing poor data utilization.[37;119] Correction for duplicate reads from amplification bias was not found. Therefore, the true quantity of useful data in many studies is unknown. Furthermore, it is unclear what the distributions of per-base sequencing depth actually were in these studies. Quantification of the sensitivity used by these studies(Table 2.1) would have helped

the experimental design for this work.

The final issue concerns the presence of background signal. Transcriptional noise is a phenomenon that is detectable with high sensitivity methods such as deep RNA sequencing.[113;114] RNA sequencing can detect low abundance signals such as residual genomic DNA,[127] low abundance transcripts,[24;26;27;128] and spurious transcription.[113;114] It is unclear how this noise was distinguished from signal in these studies, despite low and unstable depth of coverage reported by many studies.[30;31;32;121] In the most sensitive study reviewed (Table 2.1) the authors report "...less than 60% genes in the genome had their length completely covered by at least one read".[32] The effect of background signal on false positive or type I error is largely ignored in the literature.

These issues highlight the need for standards in RNA sequencing and transcriptome mapping, similar to MIAME standards for microarrays.[129] Clearly there are a number of challenges for transcriptome mapping studies, especially with unknown type I and type II error rates typical of exploratory projects.[25;26;27;28;29;123;124;127;128;130] However, by addressing technical obstacles explicitly with bioinformatic methods, the amount of useful data can be quantified. Taken together, these issues informed the experimental design and qualification of sensitivity for this project.

2.4 Lessons and Objectives

The biofuel producing bacterium *C. acetobutylicum* is an excellent platform organism for bioprocesses. A crucial challenge for productivity in this and other organisms is the tolerance of the host to large biofuel concentrations. Knowledge of the stress response systems of this organism are limited by an antiquated genome annotation[22] without transcription start sites, operon organization, or promoter signals.

These transcriptomic features are essential for understanding coexpression and regulatory networks. RNA sequencing is a powerful method used here to identify these features for future research on stress response systems and biofuel tolerance. This project aims to identify transcription start sites in the *C. acetobutylicum* genome while addressing frequently ignored issues in sequencing approaches.

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[44]. 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. This procedure allowed the experiments to be synchronized with respect to optical density (OD). 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 (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 (TE buffer) 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 was 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 were 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 were 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 terminated 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 five lanes of an Illumina HiSeq 2500 at the University of Delaware Sequencing and Genotyping Center (Newark, DE, USA).

3.4 Data Processing, Alignment

Paired-end sequencing resulted in 749,709,771 pairs of 76 bp reads which are to be deposited in the Sequence Read Archive (SRP052867). Summary statistics for the libraries are shown in Table A.1. The basic bioinformatic processing pipeline is described on Github (https://github.com/MatthewRalston/NGS_scripts) and in 4.3. In brief, the fastq headers are briefly pre-processed for downstream applications by concatenating the two columns of the Casava 1.8+ header with an underscore. Then, remaining sequencing adaptors were removed from the reads with Trimmomatic[131], an algorithm that recognizes and removes user-supplied adapter sequences. Base quality is adjusted by trimming to the minimum Phred base quality of 20, corresponding to a base-calling error probability of 0.01. Before aligning to the *Clostridium acetobutylicum* ATCC 824 genome, the data were subjected to *in silico* ribosomal RNA removal by aligning the reads to the rRNA sequences with Bowtie 2.1.0[126]. 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 before removing duplicate reads with SAMtools[132] and Picard[125]. These programs verify the integrity of the alignment file, sort and index the alignments by read name or location, and remove duplicate reads from preferential PCR-amplification.

3.5 Depth of Sequencing

Coverage vectors for each strand were initially acquired with BEDtools[133]. Coverage vectors for each transcript were then acquired with a custom Ruby script. Manipulation, summarization, and visualization of these data was performed in Julia, R[134], circos, rails, and d3. Scripts are also available on Github.

3.6 Transcriptome Assembly, Quality, and Annotation

Reference assembly was done with Trinity[24]. 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" were added to each read name in the alignment) 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. Several additions were made to this software. Specifically, unpaired reads and strand specific alignment were integrated into the coverage/alignment statistics. Additional Julia scripts were used to extract transcripts from the genome and calculate spreadsheets and summary statistics that were plotted in R and Gadfly.

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 aligned to the genome with BLAST[135] and BLAT.[136] It was determined that BLAST produced comparable and superior alignments in most cases; the BLAST alignment was used as a result. The aligned transcripts were converted to bed format, processed, converted to genePred and ultimately to

gtf format. Transcripts that completely and uniquely aligned to the genome with <30bp of gaps in the alignment were selected for further analysis. The gtf format assembly was then combined with the reference proteome for comparison.

Assembly statistics were produced with a ruby script, grouping transcripts as 'standard' (reference-ORF containing) or 'novel.' For the standard transcripts, UTR lengths and IGRs were identified and compared with both the reference annotation and according to the operon organization by Paredes *et al.*[137] Assembly quality was assessed with specific examples of canonical genes and curated through a customized genome browser. These regions were probed for agreement between known transcriptional start sites, transcript sizes, ORF boundaries, promoter, and terminator annotations. ORFs were predicted with transdecoder and subsequently annotated in RAST.

3.7 Promoter Prediction

A promoter prediction tool was developed in ruby for the detection of bipartite promoter motifs and transcription factor binding sites. Regulatory motifs were acquired through a database of *B. subtilis* transcription factor binding sites, DBTBS.[138] These motifs were then processed into their components, for example, the -10 and -35 box components of the σ_A promoter motif. Position-specific probability matrices were created from these components and used to scan the *C. acetobutylicum* genome with the MAST algorithm.[68] The results were then parsed into gtf format, filtered, and uploaded into the genome browser to facilitate assembly curation. This tool is available on Github (www.github.com/MatthewRalston/PromoterPrediction).

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

Read counts per transcript were quantified with HTSeq[139]. Raw count data were visualized and normalized in R. The data were regularized following the conservative approach of DESeq2[140],[141]. 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)[142]. Data were also processed manually for visualization in Circos graphs[143].

3.9 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. The source code of the automatic feature extraction was adjusted to be closer to the original algorithm.[144] Additional per-cluster metrics were added, described in the project README. This allowed visualization of the parameter sweep results for optimization of clustering parameters. Exploratory data analysis was done in R.

3.10 Web Content

Interactive web material was generated using a mixture of Ruby on Rails, javascript, HTML, and CSS. The d3 library[145] was used for dynamic content and interactivity. Circos was used to generate the larger circular plot visualization. These web pages are hosted on github for access by collaborators.

3.11 Genome Browser

The genome browser was designed as a web application to address issues of speed, data density, and flexibility. A simple database was created to host coverage and annotation data. No joins were necessary for data retrieval so the schema consisted simply of two tables. Simple indices were designed for each table to optimize retrieval.

The application layer was written in Ruby, utilizing the rails framework. Queries were pacified to prevent SQLi. A simple object-relational model(ORM) was used to design the interaction between the application layer and the database, although a customize query system was developed for the depth/coverage data retrieval for increased speed, bypassing the ORM and returning simple JSON-formatted text. The application layer consisted of verification protocols to ensure minimum record requirements, validity, and more.

The user interface was designed as a webpage with dynamic web content featuring the d3 library.[145] Queries are passed with simple "GET" requests, completely separating the application layer from the user. Retrieved data is passed to the user interface as JSON text and converted into SVG using javascript. The browser ui is depicted in the appendix(A.2). Users can upload and view individual annotation records, as well as upload entire annotation files in gtf format. The browser itself has a number of useful features including zooming, scaling, tooltips, and more.

Chapter 4

RNA SEQUENCING

4.1 Experimental Design

The primary objective of this research was to provide the first global experimental evidence for both canonical and novel transcripts, their boundaries, and operon structure in *C. acetobutylicum*. For this objective, a strand-specific (ss)RNA sequencing approach was superior to array based and standard RNA-seq approaches for detecting strand-specific signal at high resolution. This technique offers true strand-specific signal, typically with 1-5% background antisense signal.[36] To identify these transcripts and their features at high resolution and with true strand-specificity, this technique was selected to assess a number of experimental conditions.

A fractional-factorial experimental design was selected to best sample multiple times throughout the *C. acetobutylicum* growth curve (4.1) and in response to two fermentation products, butyrate and butanol. This organism responds to resource limitation, acid/solvent stress, and other signals by activating stress response, sporulation, and other stationary-phase systems.[6;57;146] This range of conditions was selected to view transcriptomic responses to growth stage and stress in combination for analysis with ssRNA-seq.

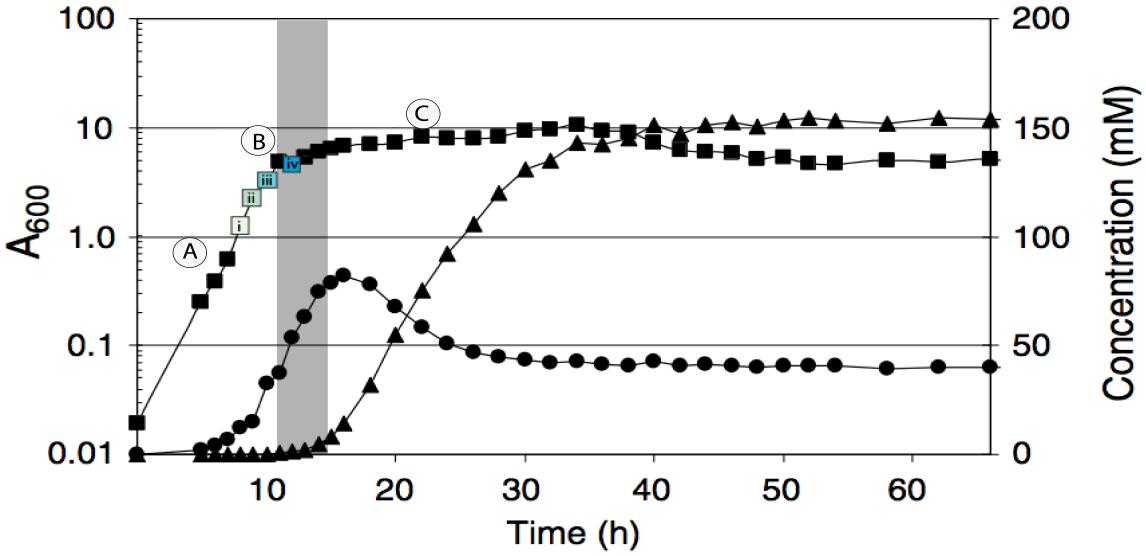


Figure 4.1: *C. acetobutylicum* Growth Curve
This growth curve, adapted from Jones *et al.*,[147] illustrates the time points selected for the experimental design. After the exponential growth phase (A), *C. acetobutylicum* cells (squares) produce carboxylic acids, such as butyric acid (circles), at increased rates during the transition phase(B). Then, the acids are reassimilated and reduced into solvents such as butanol (triangles) during the stationary phase(C). The stress types butyrate, butanol, and control were assayed in this experiment, in addition to time points 15(i.), 75(ii.), 150(iii.), and 270 minutes (iv.) after synchronization of the cultures at A_{600} of 1.0.

4.2 Desirable Data Qualities

4.2.1 Sequencing Depth

The success of transcript boundary determination with shotgun-strategy sequencing depends on two signals in the dataset: depth and complexity.[25;26;27;127;128] Here depth of coverage, sequencing depth, or simply depth is defined as the number of reads aligned to a region divided by the size of the region. In the case of a single basepair, this is simply the number of reads overlapping this base. Depth is a useful data quality for identifying absolute differences between fully transcribed

and untranscribed regions. Sequencing depth is a complex and over-dispersed random variable[140;141;148] that is non-uniform across a transcript,[130] particularly towards transcript termini.[36] Sequencing depth near transcription start sites has a complicated type II error profile that is a function of sequencing depth due to this bias. This previously discussed(2.1) and important signal is frequently used to identify expressed transcripts and their termini.

However, several factors of the experimental procedure affect sequencing depth and are not addressed in most studies. Sequence specific (hexamer,[123] GC[124] bias) or technical issues (background antisense,[36] spurious transcription[113;114]) raise variability and noise of the depth signal and have no existing bioinformatic solution. In contrast, other errors such as DNA contamination, RNA degradation, and overabundant sequences (e.g. rRNA) can be addressed with adjustment to laboratory and analytical workflows. Accounting for these issues during experimental design can improve error rates for reported sequencing depth and depth-based inferences. Specifically, the quantity of useful data in Illumina-based RNA sequencing of prokaryotes can be maximized by acquiring pure, undegraded RNA and removing ribosomal RNA transcripts. Optimal sequencing depth was the first goal for this study to improve error rates and provide a useful sensitivity metric.

4.2.2 Library Complexity

Library complexity is an additional signal that augments the information of sequencing depth. Complexity is the number of unique molecules sequenced by the experiment and can be thought of as the horizontal overlap between aligned reads.[149] Library complexity is desirable for a number of reasons, including decreased loss of sequencing depth to PCR-duplicate reads.[25;149;150] Library complexity can also be

translated directly into transcript boundaries using assembly algorithms.[24;25] Algorithmically, the assembly solution's estimates of transcript boundaries improve as both depth and complexity increase. Therefore, high depth and complexity are required for successful assembly of the dataset and determination of transcript boundaries.

Most useful assembly algorithms are overlap consensus or de Bruijn graph based, directly relying on the k-mer complexity of the dataset (where k is an integer and a k-mer is a k-length subsequence of a read) to provide significant overlaps to form the graph.[24;25] Therefore, a large amount of reads (i.e. depth) with long horizontally overlapping segments (i.e. complexity) results in a quality graph that can be traversed by an Eulerian walk. Library complexity results from the fragmentation process and the random sampling of these fragments from the library. However, complexity can be negatively affected by preferential PCR amplification of certain sequences, leading to their over-representation in the final library and dataset.[25;123;124]. Sequencing complexity is a useful data quality for both gene expression and transcriptome mapping studies. A high complexity dataset facilitates transcript boundary identification, especially in the case of low abundance transcripts, and was therefore another goal of this study.

Both depth and complexity are useful data qualities for RNA sequencing studies. Some factors such as sequence specific biases and spurious transcription are largely uncontrollable. However, other intrinsic or technical artifacts can be minimized with minor adjustments to laboratory and analytical workflows, yet they are frequently ignored in the bacterial RNA-seq studies(2.1). The poor treatment of these issues in the literature has lead to low and inconsistent coverage, sometimes with "...less than 60% genes in the genome had their length completely covered by at least one read".[32] To avoid regions of zero depth inside of annotated ORFs[32] and

false negative errors towards transcript termini, the depth-effecting factors of rRNA-removal, DNA-contamination, and over-amplificiation were optimized. The coverage/complexity signal was used to automate the inference of transcript boundaries and address false positive rates from depth-only inferences. Interestingly, coverage in turn depends on sequencing depth and the fragmentation process, the later of which can be difficult to optimize. In addition to the previously mentioned optimizations, ultra high-depth sequencing(encode ref), with hundreds of millions of non-ribosomal reads, was the appropriate method to improve both coverage and depth for this study. With these data qualities in mind, the following RNA processing workflow was established to produce an ultra high-depth sequencing dataset for transcriptome assembly.

4.3 Laboratory Workflow

A protocol was established to optimize library depth and complexity with hybridization and enzymatic steps(3.2). After each RNA manipulation step, the samples were twice washed with 70% ethanol, stored as precipitates to minimize degradation, and aliquots were taken for quality control. The quality control procedure consisted of spectrophotometric and electrophoretic analyses to ensure RNA purity and integrity. The goals and observations of the quality control process is briefly detailed first.

4.3.1 Quality Control

4.3.1.1 RNA Purity

After washing with ethanol, the absence of salts, divalent cations, and proteins was assessed through spectrophotometry. These contaminants cause RNA degradation or adversely affect the enzymatic reactions of RNA manipulation and library

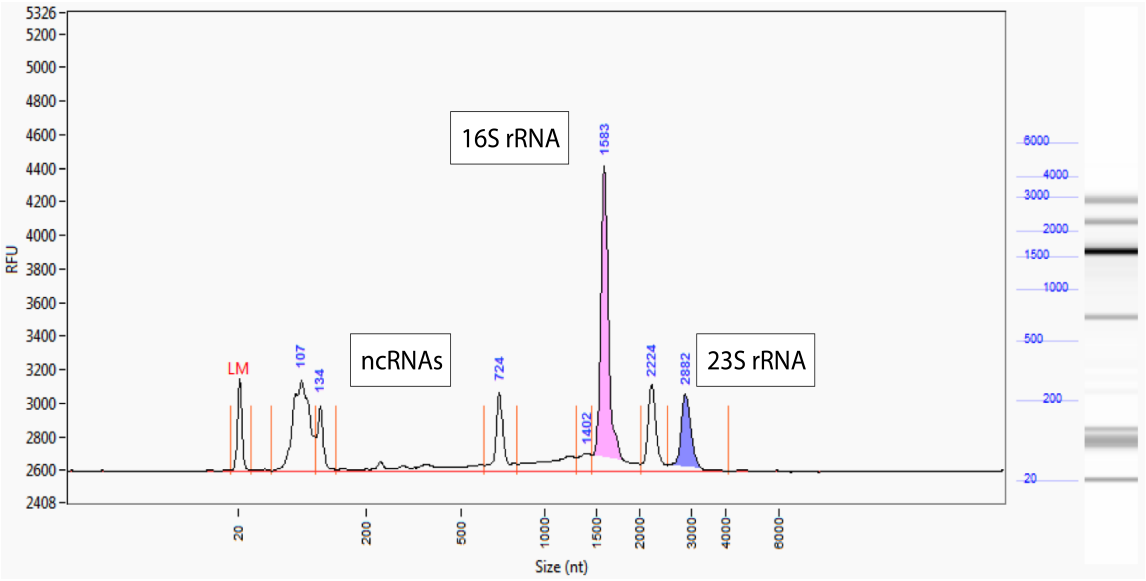


Figure 4.2: RNA Quality
In this BioAnalyzer electropherogram, sharp and intact rRNA peaks are visible along with a substantial small RNA population, resulting from the miRNeasy kit used for RNA extraction.

construction. Ratios of absorbance ($^{260\,\mathrm{nm}}/_{280\,\mathrm{nm}}$, $^{260\,\mathrm{nm}}/_{230\,\mathrm{nm}}$) are frequently used to describe the purity of nucleic acid samples, due to purine/pyrimidine absorbance maxima at 260 nm. Observed ratios of 2.0 indicated pure RNA(source), optimal for RNA integrity and library preparation. Afterwards, electrophoretic methods were used to assess ribosomal RNA removal and RNA integrity.

4.3.1.2 RNA Integrity

RNA integrity is commonly analyzed by interpreting ribosomal RNA bands obtained with electrophoretic techniques. Specifically, a small peak width of the rRNA electrophoretic bands with little background signal indicates that the RNA is high quality (e.g. RNA Integrity Number). A representative electropherogram is shown in 4.2. The results indicate that the RNA was undegraded, with sharp peaks

for the 16S and 23S rRNA bands. At each QC step (4.3), the RNA had clear pellets, clean spectrophotometric ratios, and the electrophoresis suggested that the RNA were undegraded. The passing samples were then used in subsequent hybridization and enzymatic steps.

4.3.1.3 DNA Contamination

DNA contamination was addressed in this study by DNAse digestion. This is a common step in RNA processing workflows to address residual DNA fragments. DNA and RNA have slightly different spectrophotometric properties; for example, the nucleoside Thymidine has a slightly different absorbance spectra than Uridine. Consequently, 260/280 ratios are typically higher for RNA (2.0 vs 1.8).[151] As previously mentioned, the high absorbance ratios suggested good RNA purity. It is reasonable to expect that with the treatment and these observed ratios, that the samples primarily consisted of pure, intact rRNA.

Nevertheless, the electropherograms were inspected for DNA contamination as well. Most samples displayed smooth and simple curves for the electropherogram, consistent with curves from the manufacturer showing the absence of gDNA.[152] No gDNA peaks were apparent in the RNA samples, although the gDNA fragments could have been present at a low level. Alternatively, Southern analysis of cDNA from the RNA samples, produced with and without reverse transcriptase, could have indicated the presence of gDNA. However, if there were low amounts of DNA that were not obvious from the electropherograms, they would have been somewhat uniformly distributed throughout the genome due to non-specificity DNAse treatment. This small and minor signal would be indistinguishable from background spurious transcription in the AT-rich *C. acetobutylicum* genome and would have been interpreted as background signal. In future sections, the issue of background signal will

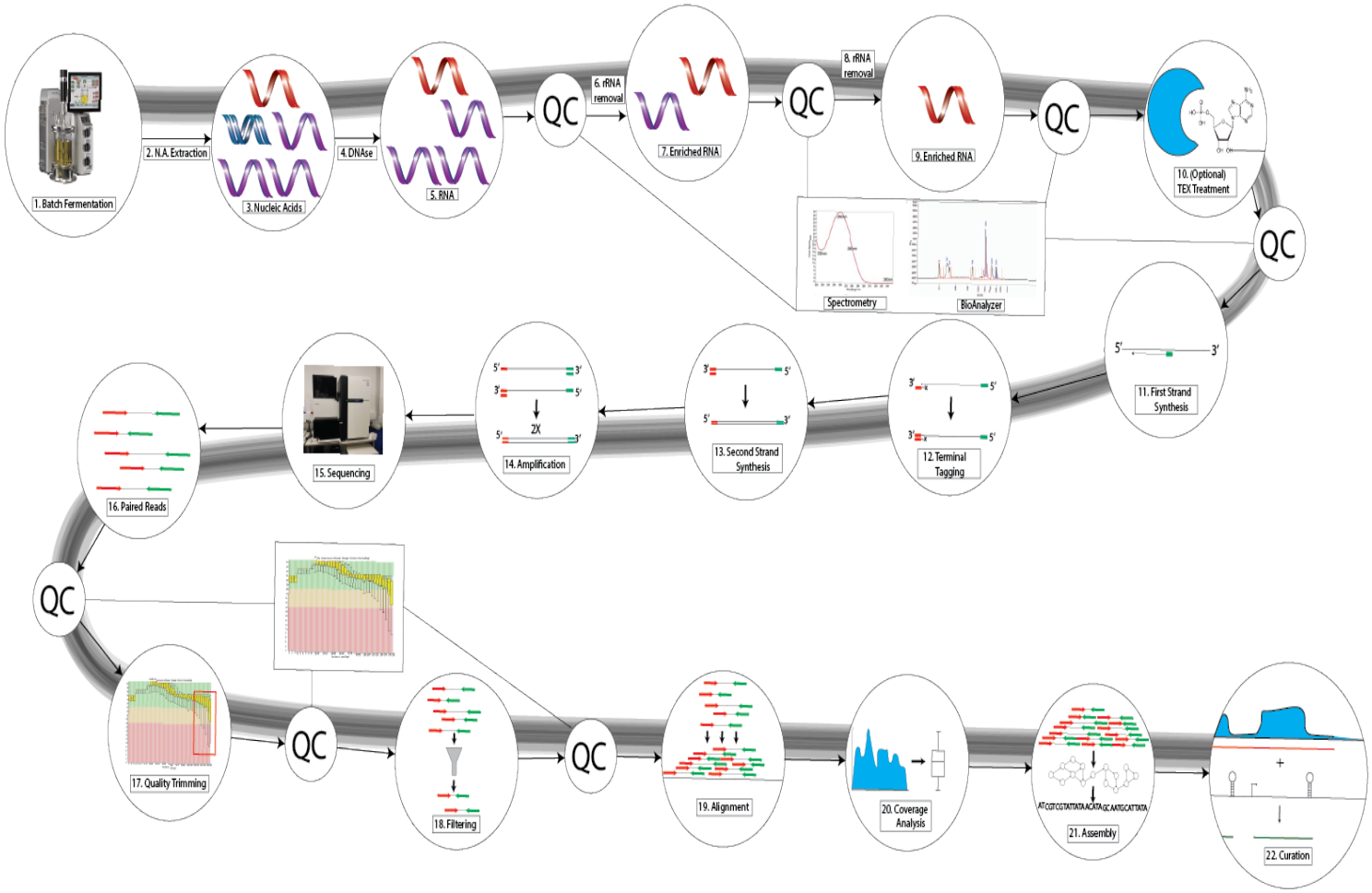


Figure 4.3: Laboratory Workflow

The laboratory workflow consisted of mRNA enrichment steps and quality control analyses to ensure RNA quality. Ligation-free library preparation resulted in paired-end Illumina reads for preprocessing, alignment, analysis, and assembly.

be discussed again as well as its remedy by distinguishing false-positive signals.

4.3.2 mRNA/sRNA Enrichment

The previous subsection discussed quality control measures that were used for initial and stepwise assessment of RNA samples. To produce a high quality sequencing dataset, RNA manipulations enriched primary transcripts (mRNAs and sRNAs) and minimized background signal (e.g. degraded transcripts, DNA) through RNA manipulations(4.3). After extraction, residual DNA was removed with DNAse treatment (Step 4., 4.3). After verifying the initial total RNA samples with the QC procedure, ribosomal RNA(rRNA) was removed with the MicrobExpress hybridization method. After additional QC, the primary transcripts were further enriched with an additional round of rRNA removal. Then selected samples - technical replicates of 2 times points(75 and 270 minutes) and all stress types (6 total) - were treated with a 5'-phosphate specific exonuclease (TEX), enriching for primary transcripts further. Primary transcripts such as mRNAs and sRNAs are produced with a 5'-NTP. Post-transcriptional processing of primary transcripts (e.g. endonucleolytic cleavage) results in 5'-monophosphate ends, which are preferred by the TEX exonuclease. The TEX treatment thus removed rRNA and degraded transcripts in these samples. This workflow maximized depth and complexity in the results by removing DNA, rRNA, and degraded transcripts. After a final QC checkpoint the enriched samples were used for library preparation(4.3) and sequencing.

4.4 Data Processing, Alignment, and Coverage Analysis

The libraries were sequenced paired-end over 5 lanes of an Illumina HiSeq 2500, producing 1.5 billion 76bp reads, averaging 25 million clusters/pairs per library(App. A.1). The reads were then processed through a customized bioinformatic

workflow (**??**). K-mer content, read length, GC-content, and additional basic qualities were assessed to ensure the quality of the unprocessed reads. Next, low-quality bases were removed to raise the quality of the sequenced bases to acceptable levels(3.4). Then, sequence reads from remaining ribosomal RNAs were removed *in silico*. After two rounds of hybridization-based removal, signal from ribosomal RNA was reduced from 95%(bacterial rRNA removal source) to 62%, representing a 7.6-fold enrichment of primary transcripts. Finally, 97% of the reads aligned to the *C. acetobutylicum* genome (4.3). Of these, 7.75M(83%) reads per sample were properly paired, that is, both mates of each pair were in the correct orientation. In total, 458,814,860 perfectly-paired non-ribosomal reads were produced and then used for subsequent analysis. This number was in excess of ENCODE recommendations for RNA-seq in the much larger human genome.[27] The remaining 17% of reads that were improperly paired were duplicate reads (32M) or discordantly (4M) or separately aligned reads (74M), as performed by Bowtie 2 (3.4).

However, the number of reads alone are a poor indicator of the sensitivity of an experiment; the distribution of these data throughout the genome is preferable, but is infrequent in similar studies(2.1). To better understand the depth of sequencing, it was desirable to determine the per-base sequencing depth throughout the genome. Two methods are frequently used to quantify and summarize depth. The first approach is referred to as "fold-coverage," calculated by summing the number of sequenced bases divided by the estimated size of the transcriptome or genome. However, the underlying assumption of a uniform distribution of reads is not valid for transcriptomic sequencing. A more precise approach is the second approach, calculating sequencing depth directly. By summing the number of reads aligned to each base, central tendency measures of the resulting distribution are precise estimates of per-base sequencing depth across the genome. No single average sequencing depth is

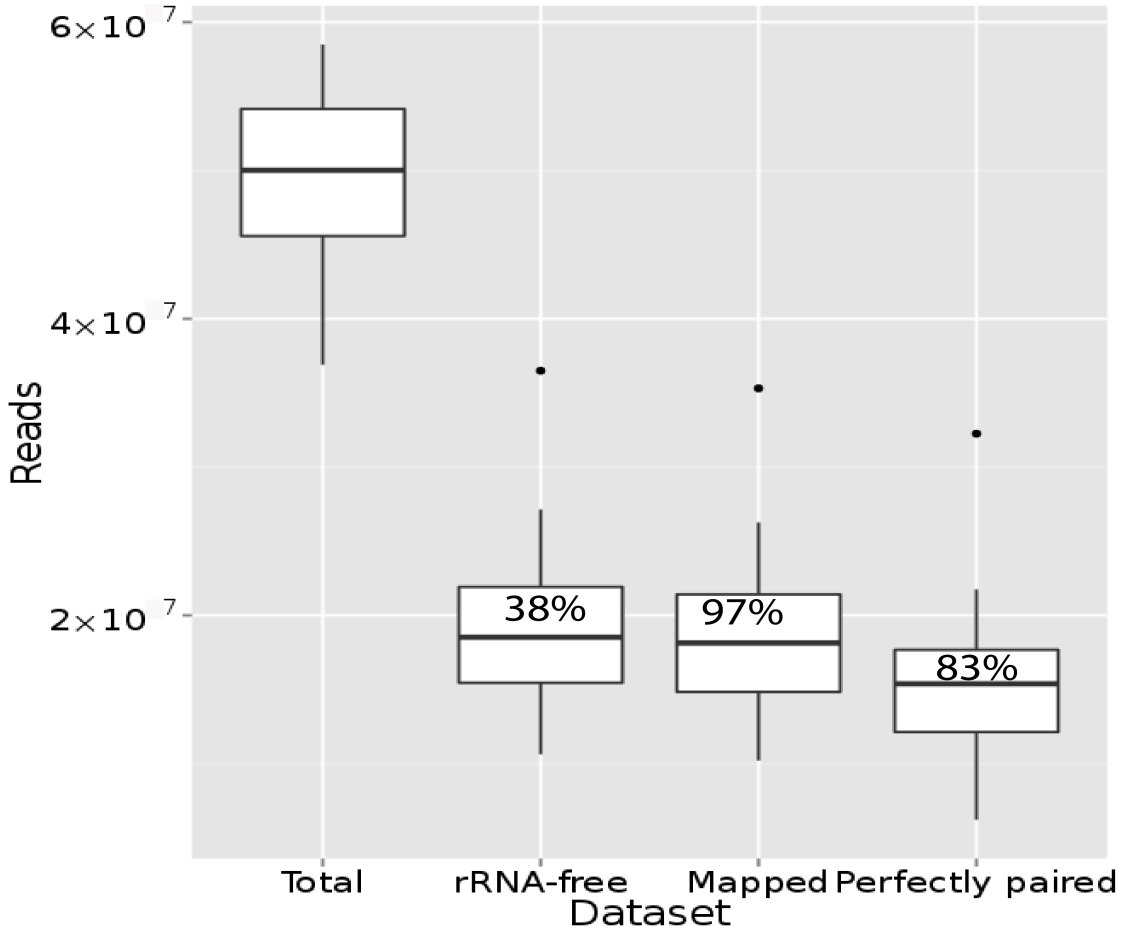


Figure 4.4: Sequence Read Processing and Alignment
An average of 50 million reads (25M clusters/pairs; left) was produced for each of the 30 libraries. Ribosomal RNA reads were then filtered and the remaining reads (middle left) were then aligned to the genome (middle right). Of these, 83% were properly paired reads (right), ideal for transcriptome assembly.

more significant than another (e.g. 10x vs 9x), although increasing depth provides additional terminal reads for transcript boundary identification.

Empirically, it seems that a coverage of 100-200 million(M) 100bp paired-end reads is sufficient to detect low abundance transcripts in the 60-140 megabase(Mb) hg19 human transcriptome,[27] although other studies claim that this number could be as high as 700M.[153] This sums to 20-40 gigabases(Gb) of sequencing, 120-660 times the conservative estimate of the size of the human transcriptome. In the case of *C. acetobutylicum*, the maximum possible size of the transcriptome is 8.2Mb, with a realistic estimate of 4-6Mb. With the 450M properly-paired reads described here, 68.7 Gb were sequenced for a much smaller transcriptome, approximately 11-17 thousand times its length. This estimate suggests that, cumulatively, this study achieves comparable or superior fold-coverage than recommended by these guidelines using the first method for depth calculation.

In terms of actual sequencing depth, however, requirements for bacterial transcriptome sequencing are unknown despite recent efforts[25;26;27;127;128] and differing views on detection limits.[127;154] In this study, the definition of ideal "coverage" is strictly a sequencing depth greater than one from one transcript boundary to another. In most species, these boundaries are unknown and their identification is complicated by uneven sequencing depth at transcript termini.[36] The discovery of transcript start and stop sites thus depends on per-base sequencing depth computed using the second method above.

A median of $>$ 10x coverage per base and per strand was observed for each of the 30 libraries (4.5). Cumulatively, the median per base coverage is 156x throughout the genome(4.6), generally considered very deep. The median depth in truly transcribed regions is greater as shown in the next chapter. To clarify, some of the depth described by this distribution(4.6) was due to previously discussed background

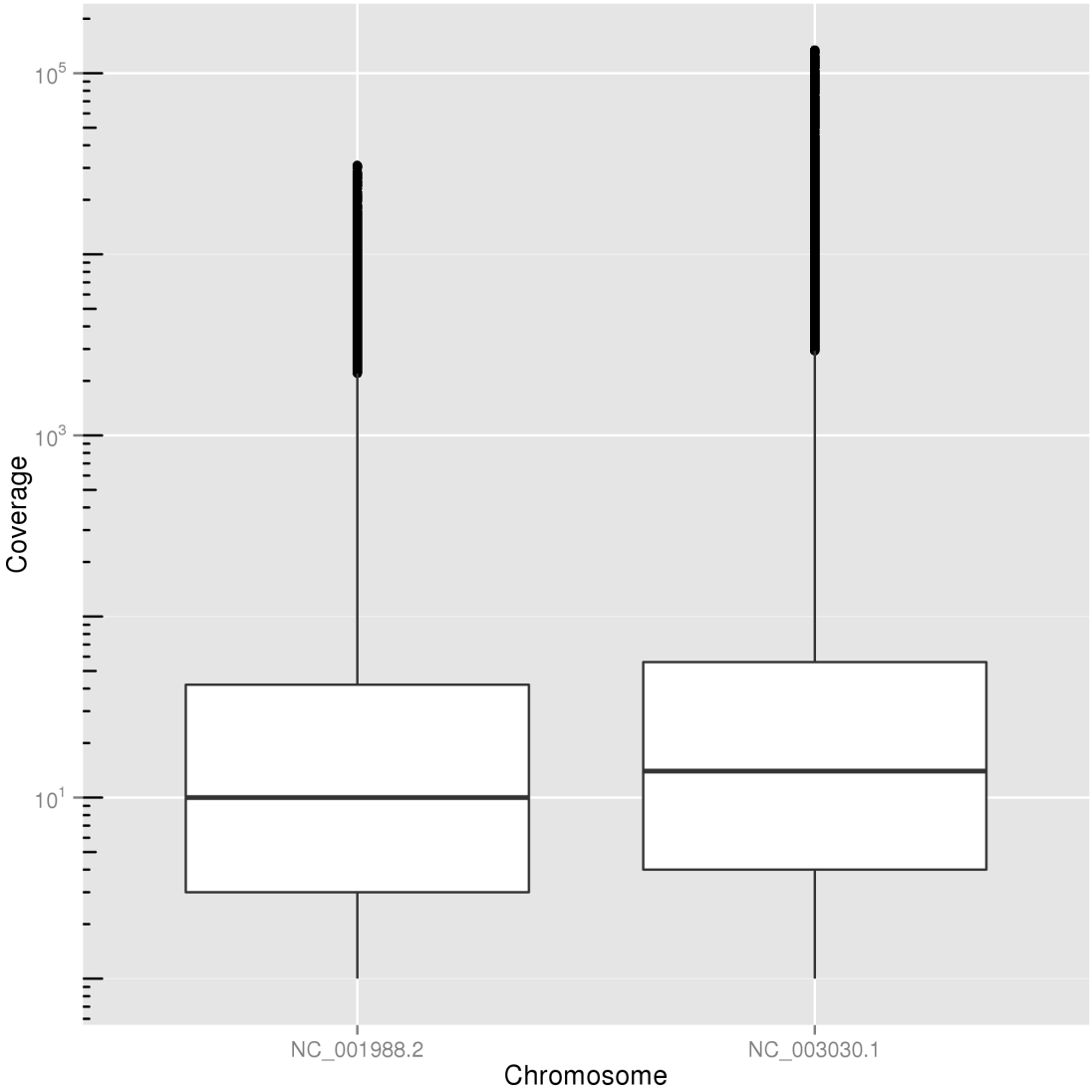


Figure 4.5: Representative Per-base Sequencing Depth
These boxplots show the distribution of per-base sequencing depth throughout the pSOL1 megaplasmid (left) and the *C. acetobutylicum* chromosome (right) for a single library. The median depth in each library was greater than 10x.

signals such as DNA contamination[127] or spurious transcription.[113;114] Background signal is indeed a pressing concern for RNA-seq,[27;127;150] complicating the determination of transcript boundaries. After describing read counts, fold-coverage, and per-base sequencing depth it is clear that this study possesses unprecedented sensitvity for transcriptome mapping.

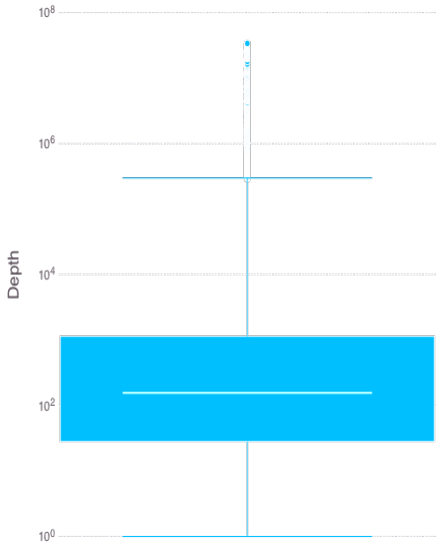


Figure 4.6: Cumulative Depth Boxplot

The distribution of per-base sequencing depth illustrates high sensitivity.

In summary, the data suggested a successful first aim for this project: a quality RNA-seq dataset for subsequent assembly and annotation. The experiment and RNA processing protocol were designed with depth and complexity in mind. Primary transcripts were enriched and contaminants were removed, controlling for RNA purity and integrity after each manipulation. Thirty samples were sequenced over six lanes, resulting in 1.5 billion reads, with 458 million properly-paired reads aligning to the genome. Analysis of the aligned sequences demonstrated consistently high primary transcript enrichment, alignment rates, and sequencing depth. This depth of signal is comparable or superior to many similar studies in prokaryotes and to guidelines for human genome sequencing.

Chapter 5

TRANSCRIPTOME ASSEMBLY

The high sequencing depth achieved in this experiment suggested that even for low abundance transcripts, boundaries would be determined effectively. To this end, transcriptome assembly was the computational technique selected for its speed and dependence on both sequencing depth and complexity.[25] Recall that assembly is resistant to the high variability of expression measurements in experiments with significant depth. It is also resistant to sequence-specific biases and low complexity background signals.

However, it is known that transcriptome assembly is sensitive to other components of sequencing datasets such as residual adapters, substitution errors, and duplicate reads.[25] In addition to the aforementioned quality trimming, adapters and primer-dimers were removed from sequenced reads and duplicate reads were removed from the dataset after alignment. This trimming and filtering strategy resulted in both paired and unpaired strand-specific reads aligning to the genome. As previously described, a large subset (450M, 83%) of the reads were uniquely aligned and properly-paired according to the forward-reverse(FR) sequencing chemistry. The remaining data consisted of 74M discordantly aligned reads. Two assemblies were conducted to understand the effect of the extra unpaired or improperly-paired reads compared to an assembly of the properly-paired data alone.

A wide range of open source tools and approaches optimized for the transcriptome have been reviewed recently.[25] Trinity was selected for transcriptome assembly

after comparison with other existing approaches. Trinity[24] is a flexible de-Bruijn graph assembler that accepts paired and unpaired reads alone or in combination. Perhaps more importantly, Trinity has strand-specific assembly options and improved support for bacterial genomes. This transcriptome assembler was optimal for this dataset and assembly optimization.

5.1 Initial Assembly

Initial transcriptome assemblies were conducted for the full dataset and the subset of properly-paired reads. Both assemblies were compared to address questions regarding assembly performance (Table 5.1). Would extra improperly-paired or unpaired reads improve the precision of boundary estimates, potentially as terminal or bridging reads? Alternatively, the properly-paired subset could have resulted in a simpler and cleaner graph for traversal by the assembly algorithm. Both of the resulting assemblies were compared by their qualities, such as assembly size, transcript lengths, and inclusion of the reference protein annotations.

In this comparison, the assemblies were inspected for errors that affect these qualities. The previous chapter described techniques for the minimization of frequently ignored background signals, not quantified by similar studies. While spectrophotometric and electrophoretic analyses suggested pure RNA, some residual signals are often encountered in RNA-seq studies.[127] Automated methods such as assembly encounter difficulty when background and overlapping signals are sufficiently complex. To identify potential false positives, the results were inspected for misassemblies, artifacts from the graphs constructed for the dataset. Documentation and analysis of these artifacts was required for assembly selection.

	All Reads	Proper Pairs
Transcripts	2874	4177
Sequenced Mb	6.1	7.2
Length Range	200-28kb	200-35kb
ORFs	2389 (63%)	3347 (89%)
Standard Transcripts	796 (28%)	1057 (25%)
Standard Mb	3.7 (61%)	4.6 (64%)
Novel Transcripts	2082 (72%)	3120 (75%)
Novel Mb	2.4 (39%)	2.6 (36%)

Table 5.1: Assembly Comparison: Proper-pairs Produce Large, Inclusive Assemblies This table contains statistics for the two transcriptome assemblies, the first with all sequenced reads and the second with only properly-paired reads. The total number of assembled transcripts and the size of their span is reported. A group of transcripts contained the majority of reference ORFs, referred to as the "standard" set of transcripts. The number and percentage of included reference ORFs are both provided. Additionally, the number of the standard transcripts and their span is provided. Finally, these statistics are also presented for novel transcripts.

5.1.1 Assembly Comparison

First, simple statistics were compiled for both assemblies (Table 5.1). The transcripts that contained reference protein annotations (referred to as "standard" transcripts), were approximately 25% by number of assembled transcripts, yet they accounted for 63% of the assembled basepairs for both datasets. Upon inspection, the assembly from the subset of properly-paired reads was larger and more inclusive, recalling 85% of the reference ORFs. Also, this assembly had higher per-base sequencing depth in both "standard" and "novel" transcripts (5.1). While the transcript lengths were comparable (5.2a), the assembly from the total amount of aligned reads had lower expression, size, and inclusiveness of the reference CDSes compared to the subset. Many factors can cause misassembly errors, not the least of which are discordant reads, which may be handled poorly by the assembly algorithms.

While the unpaired reads most likely did not negatively affect the assembly, the 74M additional discordant reads caused errors during the assembly process that lead to misassembled transcripts. Ultimately, the misassemblies decreased the total number of assembled basepairs and the inclusiveness of reference ORFs in the assembly from the total dataset. Therefore, the transcript coordinates from the uncurated properly-paired assembly were used for further analysis of feature length, UTR length, and expression.

5.1.2 Uncurated Assembly Statistics

In the uncurated assembly, 4,177 transcripts spanning 7.18Mb were assembled. This size is 88% of the maximum possible size in *C. acetobutylicum*. Each transcript aligned to a single location in the genome with >98% identity and less than 30bp of gaps, suggesting high quality assembly results. Of these, 1,029 standard transcripts spanning 4.56Mb contained 3,225(86%) reference protein annotations. The remaining 3,120 (75% by number, 36.5% by basepairs) were potentially novel transcripts, lengths ranging from 200-32.7kb. These whole-transcriptome statistics suggest that the *C. acetobutylicum* transcriptome is large and complex, in agreement with previous findings(keerthi BMC).

Additional data showed the characteristics of the transcripts themselves and painted a complicated picture. The standard transcripts, including mono and polycistronic transcripts, were larger than the novel set(5.2b). More surprisingly, they were larger on average than estimates of the mean transcript size in *E. coli*[155]. In addition, the standard set possessed higher levels of expression(5.1). Together, there was a trend between length and expression that divided the novel transcripts into distinct classes(5.3). The majority of the novel transcripts were short in length (200-500bp) with low read counts. Depending on local depth and annotation patterns,

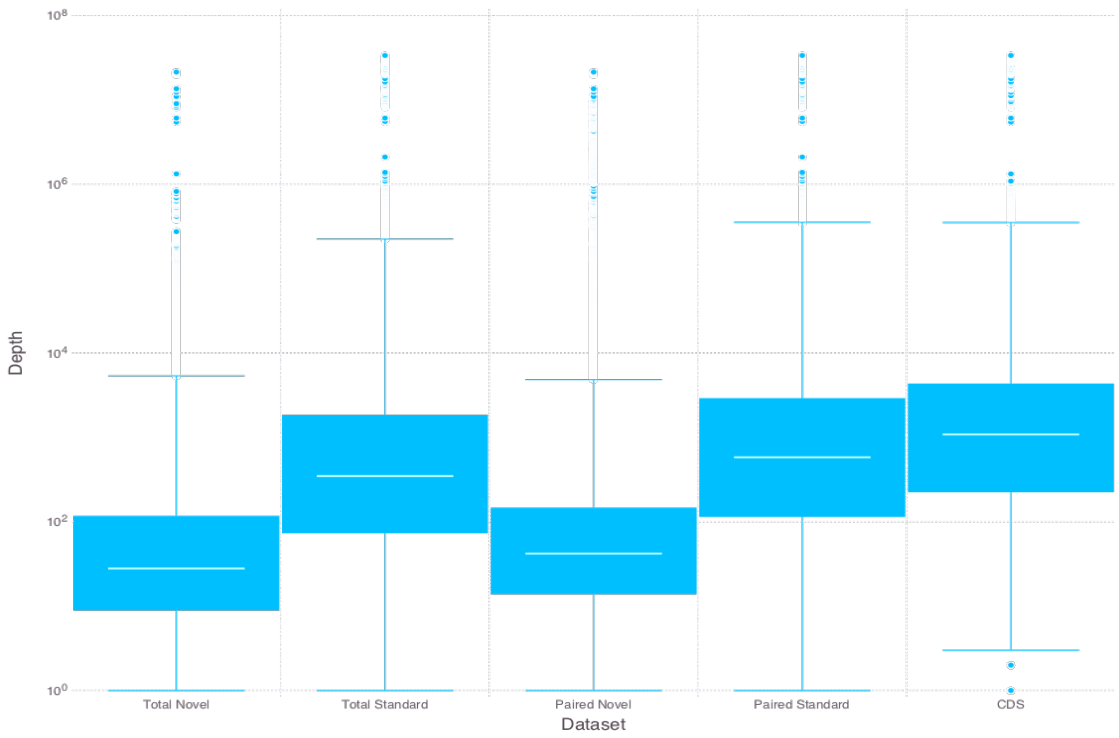
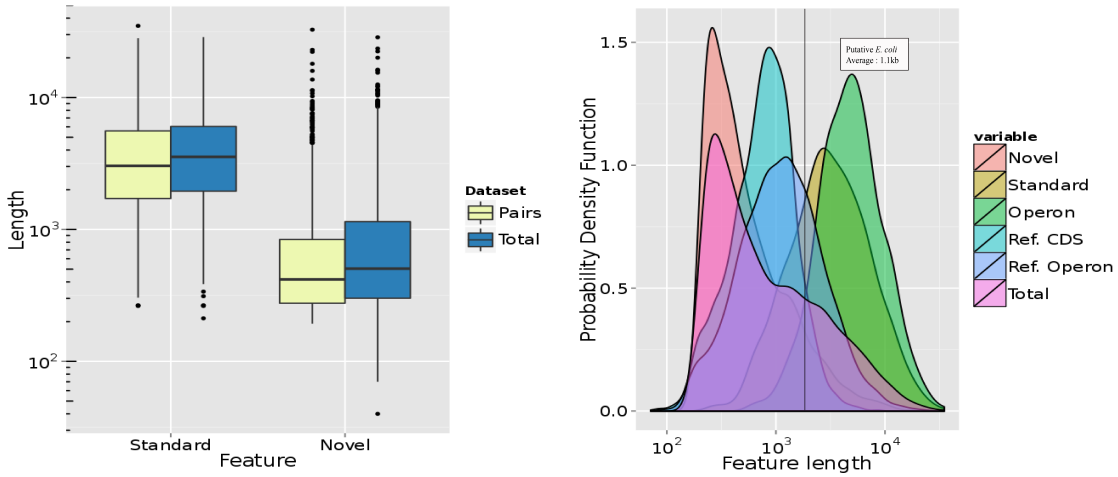


Figure 5.1: Depth Comparison: Increased Depth Observed in Properly-paired Assembly

Clearly, the standard transcripts (middle left, middle right) have higher per-base sequencing depth than novel transcripts (far-left, center). In fact the distribution of depth in standard transcripts is comparable to the reference ORFs/CDSes themselves (far right). A noticeable albeit insignificant difference can be observed between the two assemblies in term of their per-base sequencing depth. Boxplots on the left show the distribution of per-base sequencing depth from transcripts assembled from the Total dataset (left, middle left). The properly paired dataset shows a slight increase in sequencing depth for novel (center) and standard (middle right) transcripts.



(a) Length Comparison

(b) Feature Lengths

Figure 5.2: Transcript Length Comparison and Uncurated Feature Lengths
a) Length Comparison: Transcripts from the assembly of properly-paired reads (yellow) have comparable lengths compared to the assembly from the total dataset (blue).
b) Uncurated Feature Lengths: Various classes of transcripts and their associated lengths are depicted here, including polycistronic/operonic transcripts (green), all standard transcripts (yellow), novel transcripts (orange), and more. The standard transcripts appear to be slightly larger ($<$ 300bp) on average than those from *E. coli*,[155] likely suggesting misassembly.
With improved inclusiveness for reference proteins (Table 5.1), increased expression levels(Figure 5.1), and comparable transcript sizes a), the uncurated assembly from the properly-paired reads was selected for further evaluation.

some of these putative transcripts were likely technical artifacts. Longer novel transcripts with similarly low read counts were most likely assemblies of background (1-5%) or antisense signal. Outside of these groups, there were a number of highly expressed, short, novel transcripts that could reflect small peptide encoding transcripts or small RNAs. Equally expressed and larger transcripts could also represent novel transcripts and protein encoding genes. The trend between transcript length

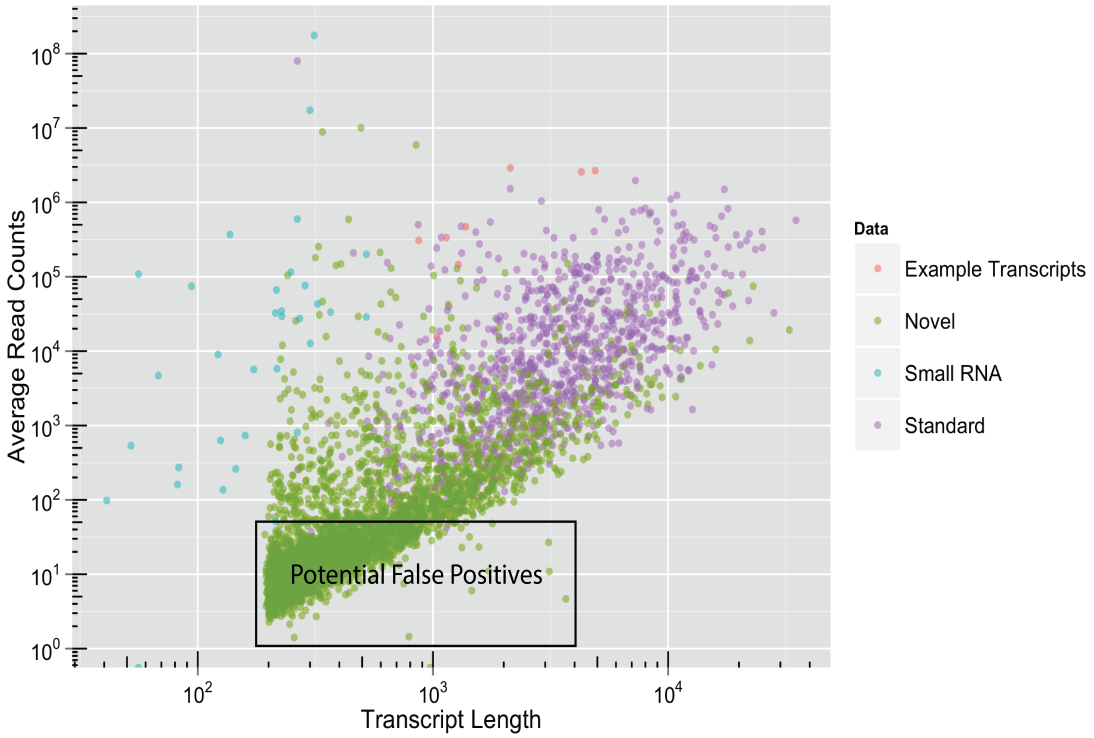
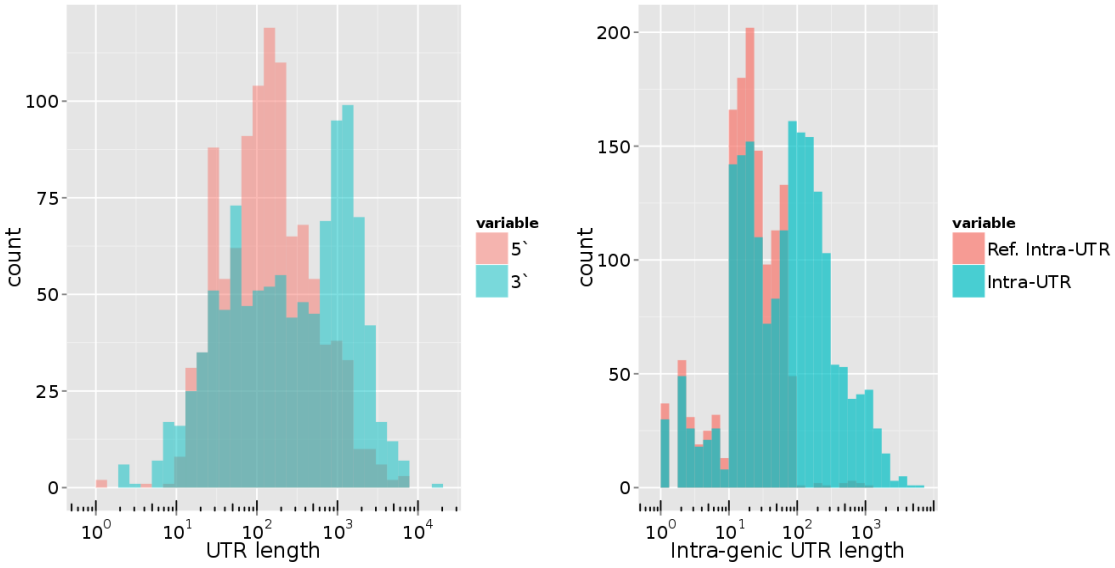


Figure 5.3: Expression (Avg. Read Count) vs Transcript Length
This scatterplot shows the standard and novel assembled transcripts along with previously verified small RNAs[44] and a few curated example transcripts that will be discussed shortly. It seems that with verified small RNAs and the standard transcripts, that an average read count threshold occurs between 50 and 100 reads. Short transcripts with low read counts could represent false positive transcripts, depending on local background sequencing depth. A large number of transcripts possess comparable lengths and read counts to standard transcripts. This suggests that despite the false-positive signal captured with this level of sensitivity, truly novel transcripts were detected.

and expression indicated the presence of both novel transcripts and technical artifacts in the assembly results, suggesting that further investigation and correction would be necessary.

An additional illustration of misassembly was seen in the distribution of untranslated region (UTR) lengths (5.4a). A number of the standard transcripts possess



(a) 5' and 3' Untranslated Regions

(b) Intra-operonic Untranslated Region

Figure 5.4: Untranslated Regions

a) While terminal UTRs can contain regulatory sequences, most in *E. coli*[156] are around or less than 100bp. It is exciting to speculate about the existence of unannotated proteins, although these results most likely indicate misassembly.

b) Many intra-operonic UTRs agree with work from prior predictions[137], although misassembly has not been excluded for the UTRs described here. Curation of the initial assembly could reveal unannotated proteins within these large UTRs.

5' and 3' UTRs that were several hundreds of basepairs in length, while most UTRs previously determined in *C. acetobutylicum*[48;49;55;72;157] and *E. coli*[156] are approximately 100bp. Some of these could have contained riboswitches or unannotated proteins, although likely not at the frequency shown by this histogram. Therefore, it was desirable to address these misassemblies through a curation process.

Nevertheless, encouraging results were obtained from examination of the uncurated assembly of the properly paired reads. This subset produced a large number of transcripts spanning 88% of the bases of the genome and contained the majority of the reference protein annotations. The large number of assembled basepairs suggested both sufficiently high sensitivity (low false negative rate)and good k-mer complexity in the data. This truly diverse library was likely to contain rare and novel transcripts. Analysis of the novel transcript size and expression suggests that small RNAs and larger protein-encoding messages have been acquired in this dataset in addition to technical artifacts. As expected, false positive transcripts were assembled from background antisense signal or spurious transcription. Additional evidence for these background signals were apparent in large UTR lengths of the standard transcripts. After seeing evidence of these issues in both standard and novel transcripts, it was desirable to closely examine and illustrate these examples. To investigate these issues, a customized genome browser was developed as a tool for curation to increase the precision and accuracy of the transcript coordinates. The integrated curation method involving this tool is discussed next.

5.2 Exploratory Tools

As described in the previous section, inspection and description of the background signal was required to identify the previously mentioned misassemblies. Such

illustration is typically accomplished with a genome browser. Genome browsers allow the exploration of sequencing datasets in high detail, producing publication-ready images of depth, coverage, features, and more. To facilitate this exploration, flexibility was a key aspect for selecting a genome browser for both identification and correction of assembly errors.

Specifically, an ideal genome browser would display depth, coverage, and annotation data from both strands separately. Visualization of multiple coverage vectors (e.g. different conditions) in a single track has improved data density compared to multi-track browser designs. Unfortunately, "reducer" functions (max, sum, average) are not simple to compute for multiple large alignment files and existing genome browsers. In fact, many conventional browsers are sluggish to even load such large datasets.[158;159] These genome browsers did not meet the requirements for this project. Instead, a customized genome browser was constructed with flexibility, speed, and simplicity in mind to facilitate assembly visualization and curation with the sequencing data and genome annotations.

5.2.1 Genome Browser

In this genome browser, only the coverage vector was required for visualization and not the inspection of individual reads. A total of 169 gigabytes of aligned reads were summarized by 6.8 gigabytes of coverage vectors, a dramatic reduction of resource requirements. Still, these data were too large to transmit to users or perform reducing functions upon. The appropriate format for visualization and distribution of these data to the *Clostridia* research community was a web application with a database. The objective for this genome browser was simple: allow users to upload and view feature annotations (e.g. sRNAs, proteins) alongside condition-specific

coverage vectors from this sequencing dataset. The details of its construction have been described in the Methods chapter (A.2).

The finished product was a modern genome browser with an intuitive user interface(A.2). A simple database was required to host coverage and annotation records. An object relational model was required to retrieve and pass data to the web application layer for conversion into scalable vector graphics (SVG). This creation had both speed, with optimized SQL queries, and flexibility, with interactive zooming, filtering, and tooltip details. Strand specific coverage and annotations were displayed in a publication ready form. This genome browser was a tool to integrate annotation types including the transcriptome assembly, reference CDSes, and more, contextualizing these genomic features with expression data. After construction, this genome browser was loaded with genome annotations to facilitate error correction.

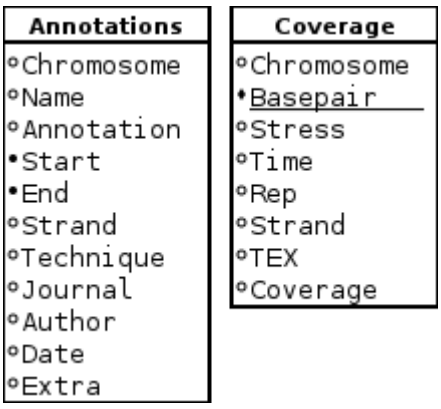


Figure 5.5: Database Tables
A simple database was designed, indexing the annotations and coverage entities on their genomic coordinates.

5.2.2 Promoter Prediction Tool

To aid the curation process, it was desirable to integrate additional annotations such as promoter and terminator predictions. For example, promoter predictions help resolve misassembly near transcription start sites. Promoter motifs are genomic signals that should correlate with expression levels at a rate predictable from sequence similarity to a consensus motif. Unfortunately, promoter annotations do not exist for *C. acetobutylicum*. After observing the extended transcripts and

UTRs in the previous section, a promoter prediction tool was developed to address these errors.

A promoter prediction tool was developed to utilize consensus sequences from *B. subtilis*[138] to predict promoter motifs in this *C. acetobutylicum*. The promoter prediction tool, described in 3.7, converts consensus sequences into models suitable for input into the MAST algorithm.[68] Consensus sequences from DBTBS[138] were used to generate models of promoter elements and transcription factor binding sites. Then, these were used to scan the *C. acetobutylicum* genome. Predictions with p $<$0.01 were then converted to GTF format and uploaded into the browser. This browser, loaded with annotations, was then used to visualize errors in the uncurated assembly and discussed in the next section.

These exploratory tools were created to improve the precision of the transcript boundary estimates by contextualizing the coverage patterns with annotations (e.g. Rho-independent terminators, promoters motifs) that explain drops in depth. These genomic signals provide biological mechanisms for the depth and coverage observations, increasing the amount of useful signals present at transcript boundaries. These annotations were combined in a customized genome browser to be used for integrative analyses of these genomic signals and sequencing data.

5.2.3 Background Signal

In the previous section, a genome browser and promoter prediction tool were described to facilitate an integrative analysis and curation

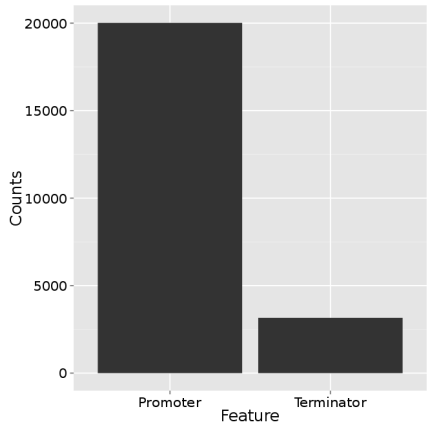


Figure 5.6: Feature Frequency Barplot of total numbers of predicted σ_A-promoters and Rho-independent terminators,

of the *C. acetobutylicum* genome. In this section, these tools were used to assess the background signals responsible for misassembly, produced from the high sensitivity of this experiment. A large number of Sigma A promoters were predicted ($p < 0.01$) throughout the *C. acetobutylicum* genome (5.6), close to 3x the number of predicted terminators. These promoters were uniformly distributed and perhaps surprisingly, were not necessarily concentrated at the beginning of transcripts (5.7). Many of these predictions were weak matches to the consensus motifs ($p > 0.001$) and would in turn have had only weak affinity for σ-factors and residual transcriptional activity. The AT-rich genome of *C. acetobutylicum* leads to an abundance of putative promoter sequences, contributing to the background signal observed both statistically and through specific examples (5.8). In addition to the previously described benefit of integrating promoter motifs into the curation process, this enabled the quantification of their prevalence and quality. The abundance of σ_A motifs in this AT rich genome could be partially responsible for some of the background signal and misassemblies.

The transcriptome assembly was inspected with the customized genome browser to understand this background signal and determine if the transcript boundary estimates could be improved by curation. The assembled transcripts matched the sequencing depth well, despite the misassemblies that were apparent upon examination (5.8). Assembly extended through some troughs in depth (transcript fusion) and beyond expression termini (transcript extension). The extent of background

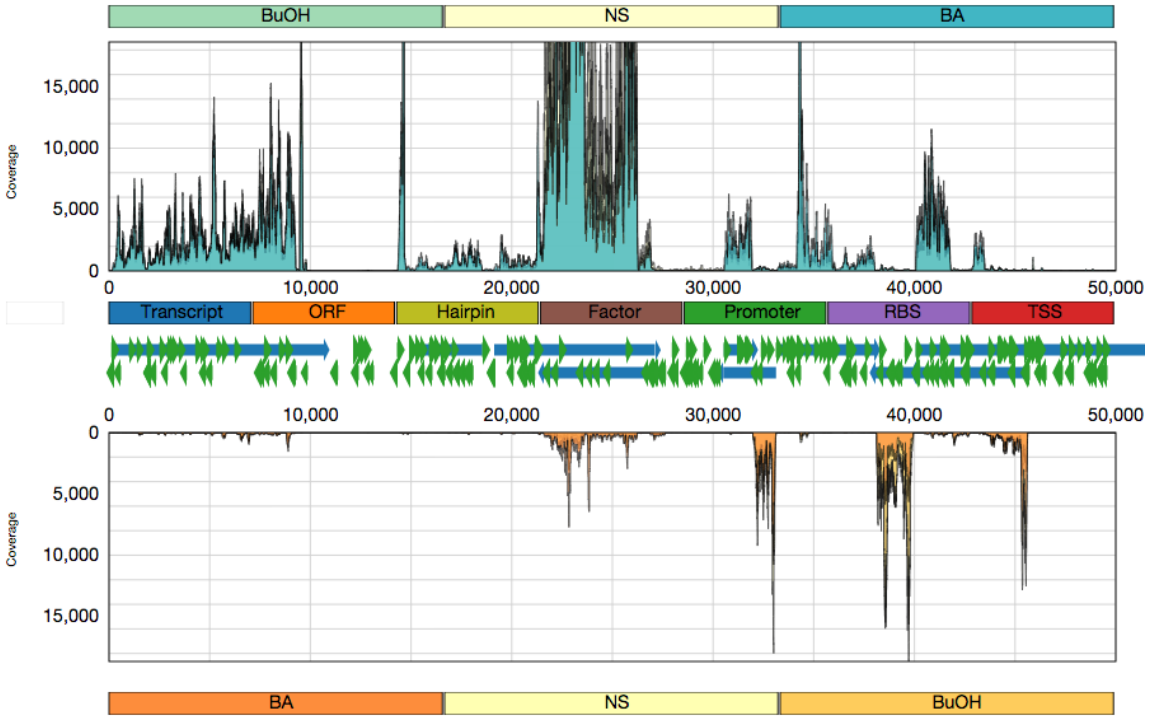


Figure 5.7: Promoter Prevalence

A representative region of the *C. acetobutylicum* genome with a large number of σ_A promoter motifs relative to the Rho-independent terminators. The high frequency of these promoters in the AT-rich *C. acetobutylicum* genome may contribute to the high background signal.

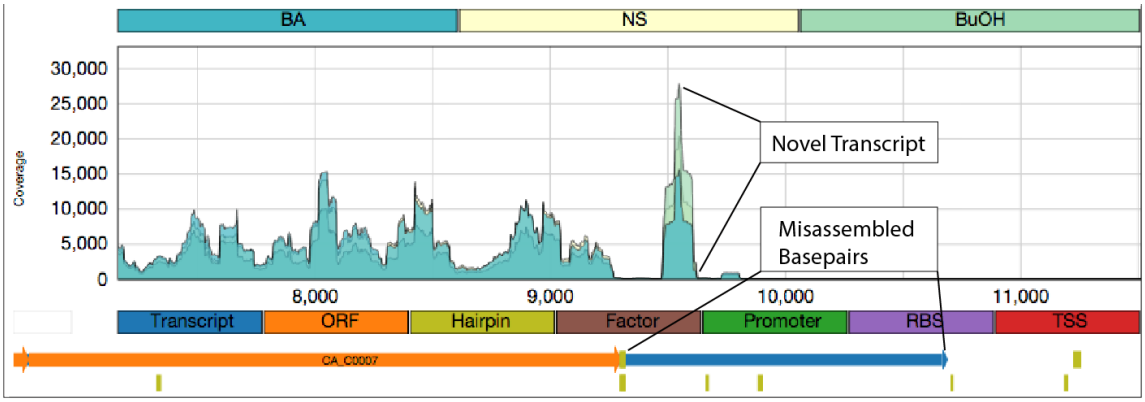


Figure 5.8: Background Signal

An example of a misassembled trancript is shown, extended beyond the hairpin/terminator into the intergenic region. A novel transcript is indicated that may have contributed to this misassembly, although the misassembly extends into obvious troughs of sequencing depth. The misassembled region and novel transcript are indicated in the figure.

signal - residual depth in seemingly inactive regions of the genome - is impossible to quantify without first distinguishing true signal from the noise through assembly curation. However, background signal is not uncommon with RNA-seq,[27;112;113;127] despite neglect in comparable bacterial studies.

Potential sources for this noise include residual antisense signal (1-5%), contaminating DNA, and spurious transcription. These signals that were minimized (ref methods) but are difficult to eliminate completely in RNA sequencing experiments(ref background signal papers). Residual antisense signal was not overly abundant(1-5%), mostly a factor of the library preparation method.[36] Contaminating DNA was minimized and not observed during quality control checkpoints. While these residual contaminants might have contributed towards the background signal, they were expected to be distributed uniformly throughout the genome. These signals could have contributed to the misassembly.

The last source, spurious transcription, can be minimized through certain

extraction and size selection techniques during library preparation. However, this experiment was designed to identify all coding and noncoding primary transcripts. The RNA extraction technique used did not exclude short transcripts, such as those from non-specific transcription. The previously described sequencing depth suggested that some of this signal was expected.[27;112;113;127] Spurious transcription was also supported by the promoter prediction frequency (5.6) and uniform distribution of these motifs in both transcribed and untranscribed regions of the genome (5.7). Similarly to the other noise factors, a small and uniform noise is expected given the extreme sequencing depth, and this noise is a likely cause of misassembly.

Fortunately, there were distinct depth patterns (5.8) that agree with promoter and terminator annotations. The separation of true signal from noise was possible by integrating these multiple datasets through the genome browser. Next, the curation process is detailed for specific examples, where previous gene-specific experiments have produced transcript boundaries. The integrated analysis corrected for these background signals, revealing precise and accurate estimates of transcript boundaries.

5.3 Example Transcripts and Curation Process

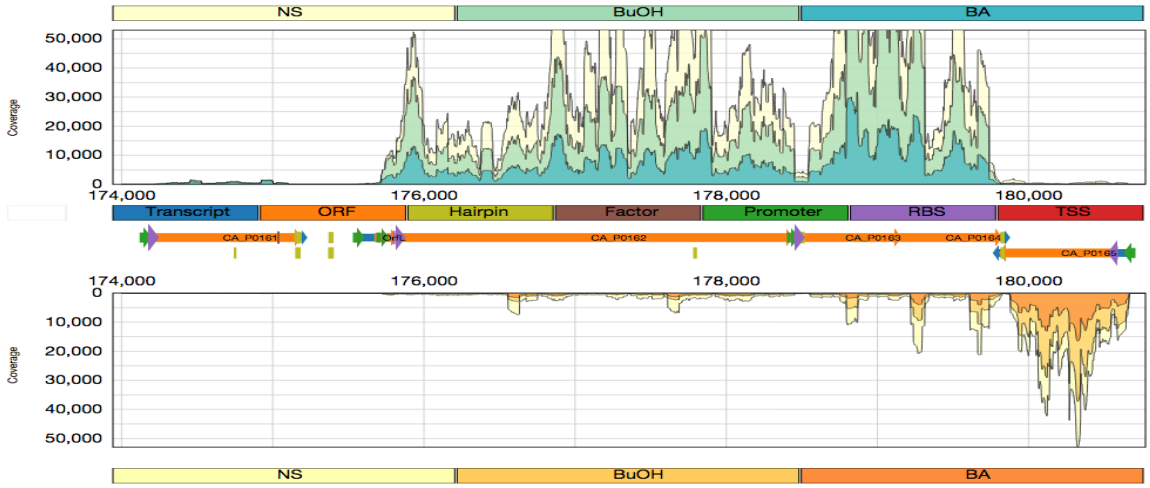
Previous sections described global indicators of misassemblies cause by background signals and the tools required to identify and address them. Correction also required clear description of these errors and a defined curation method, facilitated by the genome browser and genomic signals. These errors, not discussed or documented in similar studies, were best described for genes that have previously determined transcript boundaries. This external validation allowed true understanding of the type I and type II errors in the uncurated assembly. Six issues, listed below, were considered for each example to better understand the quality of the assembly and the degree of curation required.

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

These questions were considered for each of 5 loci where there is *a priori* knowledge. In the following examples, the uncurated assembly results were compared to the promoter, terminator, and sequencing data for the region. Misassemblies were documented when the assembled transcript disagreed with obvious patterns in sequencing depth, promoter, and terminator annotations. These examples illustrated the types of errors found in the data and details the simple methods to correct them. The first example, the Sol locus, contains 3 transcripts that display minor extension misassemblies with clear and simple solutions.

5.3.1 Sol Locus

The *sol* locus is a 7kb region on the pSOL1 megaplasmid surrounding the *sol* operon(4.3kb, basepairs 175,564-179,841). This region is responsible for the production of several solvents[47;48] and is immensely important to the physiology of the *C. acetobutylicum* ATCC 824 strain. This region encodes several enzymes including a tri-functional NAD(H^+)-dependent alcohol/aldehyde dehydrogenase (AdhE1)[47], two subunits of coenzyme-A transferases (CtfA/B)[50], and an acetoacetate decarboxylase (Adc)[49;50;160]. The region is also home to a protein SolR, which includes a helix-turn-helix motif and is thought to regulate solventogenesis[161]. These genes are vital for



(a) Sol locus

Figure 5.9: Sol Locus Overview

This operon (upper track) consists of orfL, alcohol dehydrogenase (adhE1), and Co-A transferases A and B (ctfA,ctfB). solR (far left) and acetoacetate decarboxylase (adc; lower track, right) are 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.

carboxylic acid reuptake and conversion into alcohols, a vital part of this organism's metabolism and the solventogenesis process.

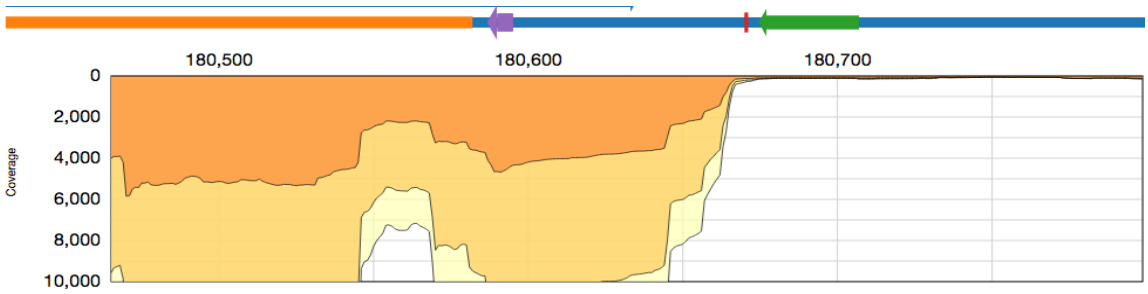
5.3.1.1 Acetoacetate Decarboxylase Transcript

In the early 1990s, several articles were published about the *sol* locus including the cloning and sequencing of adc and the *sol* locus[47;48;49;50;160]. An early study of the *sol* operon probed the adc locus, reporting two transcript sizes of 670 and 865 with Northern blot[160]. The authors also reported the major transcription start site of adc at base 180,671 of the pSOL1 plasmid. To examine the quality of our data and raw assembly, we examined this locus to observe the transcript size and locate

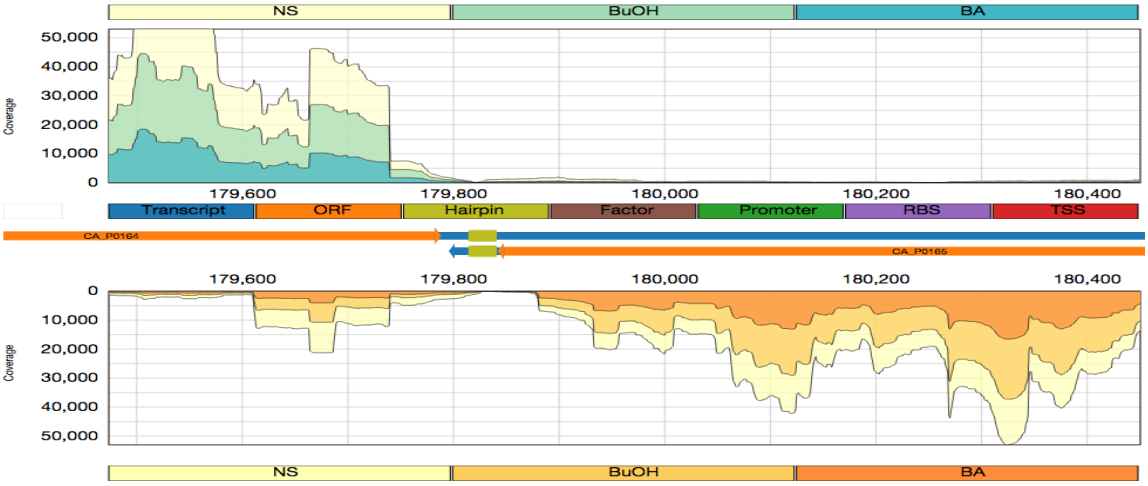
the transcription start site in our data.

In 5.10a we see the transcription start site reported by Durre *et al.*(red)[160] located very near a sustained increase in sequencing coverage just downstream of a canonical promoter motif. This pattern of coverage (cumulatively >10,000x) is sustained until a bidirectional Rho-independent terminator (5.10b). In this instance, the precise transcription start site was not estimated precisely by the uncurated assembly. The reported transcript continues for several hundred basepairs upstream of the adc TSS, despite the decrease in coverage. This artifact is most likely due to sufficient k-mer complexity in the reads mapping upstream of the TSS for the assembly algorithm to fuse these reads to the adc transcript. While this complexity is generally a good sign for the quality of this dataset, in this case a misassembly was the result. Correcting for this error (5.10c), the full transcript size is 857bp.

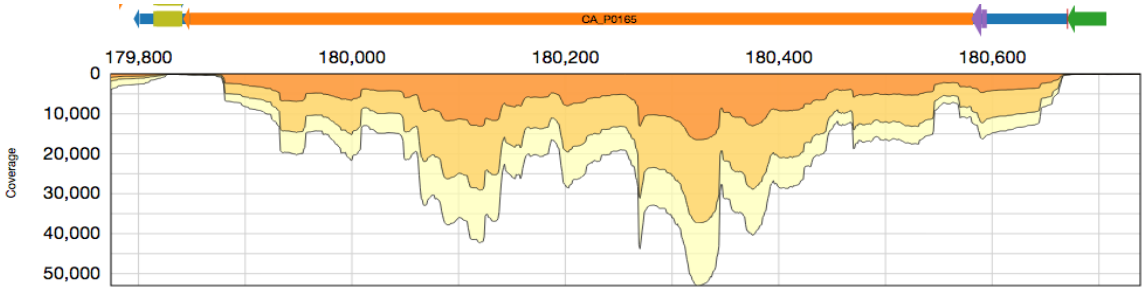
It was claimed that a 670bp product was most likely a specific degradation product or the result of a secondary transcriptional start site[160]. To investigate this, a transcript of this size would correspond to a transcriptional start site at approximately base 180,484. Unfortunately, none of the promoter motifs in the region could explain a transcript of this size in vegetative cells. After curation based on the coverage pattern, promoter and terminator motifs in this region, the transcription start site and transcript size for adc accurately match previous results. The uncurated assembly predicted a transcription stop site with good precision, in contrast to the start site. The degree of k-mer complexity, despite the low coverage in the background signal upstream of adc, suggests that library complexity is reasonably high. In this instance, the background signal poses and obstacle for automated assembly, but suggests that sufficient depth is achieved in this experiment. This type of misassembly will be referred from here on as an "extension". From 5.10b, another extension misassembly is present in our next example transcript, the *sol* operon.



(a) adc transcription initiation region on the Crick strand.



(b) Bifunctional Rho-independent terminator for *sol* operon (upper track, left) and Adc transcripts



(c) Curated adc locus

Figure 5.10: Adc locus

a) 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. b) A bifunctional terminator is responsible for transcriptional termination of both Adc and the *sol* operon. c) With minor curation, the region matches previous results faithfully.

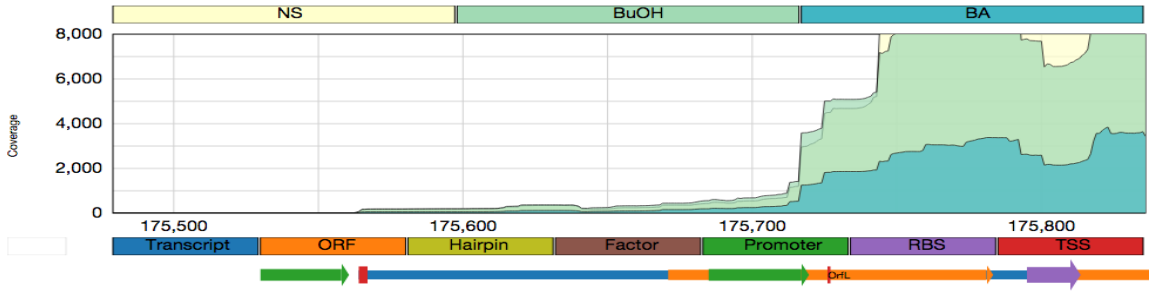
5.3.1.2 Sol Operon

In[160], the *sol* operon was investigated using a probe specific for ctfB and a transcript of size of 4.1kb was reported. Unfortunately, no blots were included as figures in this work. In[48] adhE1-based probes revealed a nearly identical transcript size of 4.1-4.2kb. Interestingly, the *sol* operon has both proximal and distal transcription start sites at 175,726 and 175,564, respectively[47;48]. Ribosome binding sites have been identified upstream of adhE1, ctfA, and ctfB[48]. From these previous studies, good recall of the transcription start sites was expected.

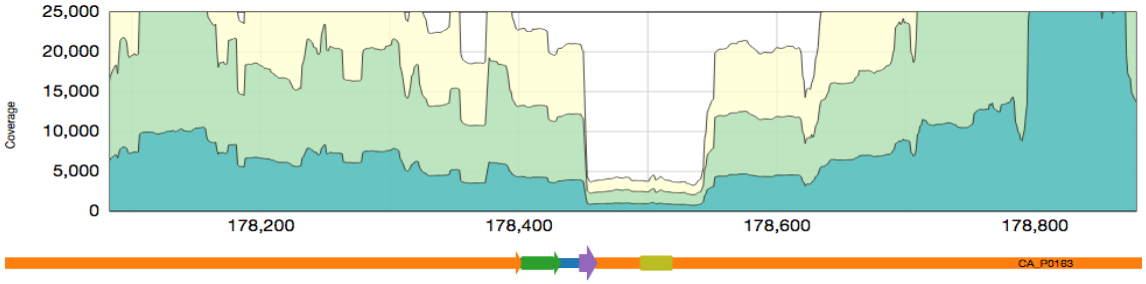
The expression level of the *sol* operon is substantial, also upwards of 10,000x coverage cumulatively. The distal transcription start site is matched perfectly (5.11a), demonstrating the precision of the assembly technique in the absence of background or residual signal. An increase in coverage is observed immediately following the proximal promoter (5.11a) in close agreement with the previous determinations[47;48]. However, the transcription stop site was not precisely determined by the assembly, owing in part to basal antisense signal from the Adc gene (5.10b). After adjustment, the transcript sizes are 4,115 and 4,277, respectively, in close agreement with the reported transcript sizes[48;160].

5.3.1.2.1 Multiple Transcripts from Sol Operon

While the results from this region agree as a whole, there is an interesting pattern in coverage in the *sol* operon near the C-terminus of adhE1(5.11b). This 100bp region is expressed at a statistically lower level (K.S.-test, $p < 0.05$) than the rest of the *sol* operon but has a standard GC content of 35%. It is unlikely that the low sequencing depth is caused by sequence specific biases, described earlier. Upon further examination of this region, we find a Rho-independent terminator with a ΔG of -9.6 kcal/mol that is not as strong as the -11.5 kcal/mol bifunctional



(a) *sol* operon transcription initiation region. The distal (left) and proximal (right) transcription start sites (red) are shown for adhE1 (far right, orange).



(b) Putative adhE1 (left) terminator, ctfA (right) promoter

Figure 5.11: Sol Operon

a) *sol* operon (orfL, center; adhE1 right) transcription start sites. The coverage and assembly data have strong agreement with previously described proximal and distal promoters and transcription start sites. b) 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 these observations.

terminator at the end of the *sol* operon. The region is also near a σ_A promoter motif of TTCATA(13)TATAAT located upstream of the previously mentioned RBS.

As mentioned above, no Northern blots figures were included in the only study, to the best of my knowledge, that uses ctfA or ctfB specific probes[160]. Most studies of the *sol* operon in *C. acetobutylicum* use adhE1-specific probes or larger restriction digestion probes[48;51;162]. One of these displays a Northern with a weak but distinct band for a 2.6kb transcript under solventogenic conditions[157]. If adhE1 were transcribed both in the classical *sol* operon and as a separate transcript, the length would be 2,716bp from the proximal transcription start site, similar in size to the observed band[157].

This region shows steep changes in coverage that are consistent across all replicates. It is likely that this observation does not merely represent difficulty sequencing secondary structure in the *sol* operon transcript. In addition to the full *sol* operon transcript, additional transcripts from the adhE1 and ctfA/B genes may be produced. Recent Northern blots from our lab seem to corroborate this finding (Alex Jones, personal communication, December 9, 2014).

In addition to correcting misassembly, identifying overlapping transcripts and alternative transcription start sites is a benefit of curation. The uncurated Sol transcript has basepair-level resolution of the distal transcription start site. With minor curation, two additional transcripts were discovered from this region and the stop site was easliy determined. For the first two example transcripts, boundaries were determined with high precision and accuracy, although it should be noted that these regions had strong expression. In the next example, high precision is achieved again, even in a transcript with lower coverage.

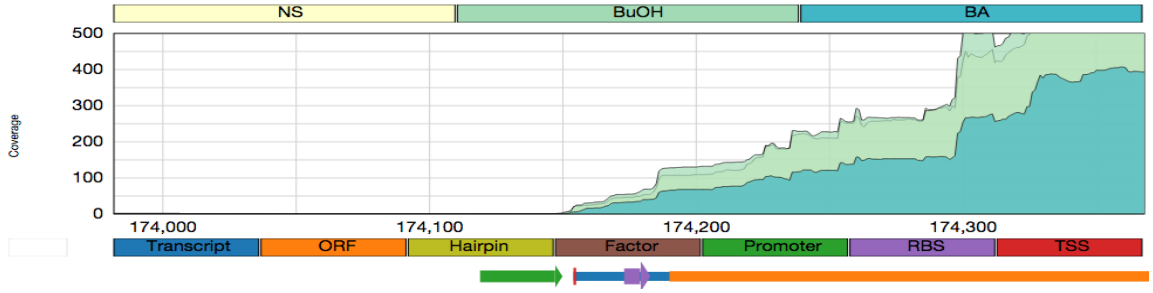
5.3.1.3 SolR Transcript

The last transcript produced from the *sol* locus considered here is from the solR gene. This gene produces two transcripts, one 1kb and the other 1.3kb[48]. A later study revealed the role of SolR as a repressor for the *sol* operon[157]. This study also produced a single transcription start site at 174,154 on the pSOL1 plasmid. As a result of examining this dataset, perfect recapitulation of the transcription start site was achieved(5.12a). The transcript from the raw assembly is approximately 1.2kb(5.12b), showing only residual coverage (cumulatively <5x) after the first terminator(5.12c). After curation (5.12d), the transcript size is 1,036bp. There was insufficient coverage in the conditions studied to strongly support the larger transcript. No alternative transcription start sites were apparent.

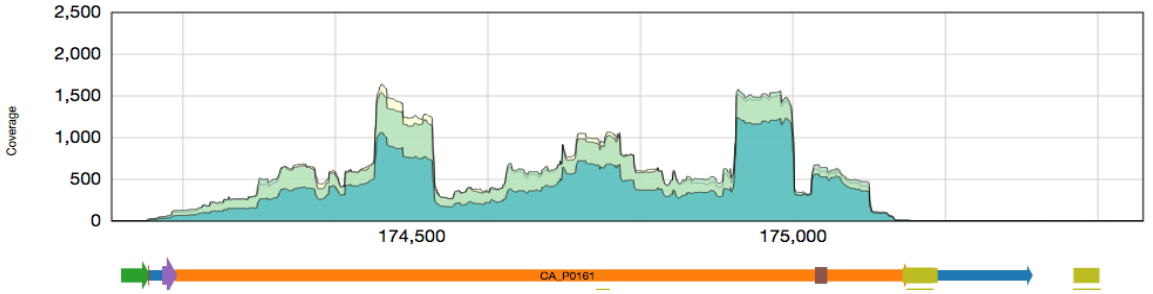
In the first locus, boundaries were accurately determined after minor curation. Examples of the first type of misassembly, the "extension" were presented. The level of complexity of the reads mapping to these regions was high, given complete assembly of these regions including low coverage background signal. Basepair level resolution of the transcription start sites was observed for 2 of the 3 transcription start sites without curation. With the misassemblies that were described here, minor curation was required to accurately determine the remaining boundaries. In the next example locus, two butanol dehydrogenase transcripts will be examined, along with the next type of misassembly that this curation approach can address.

5.3.2 Bdh Locus

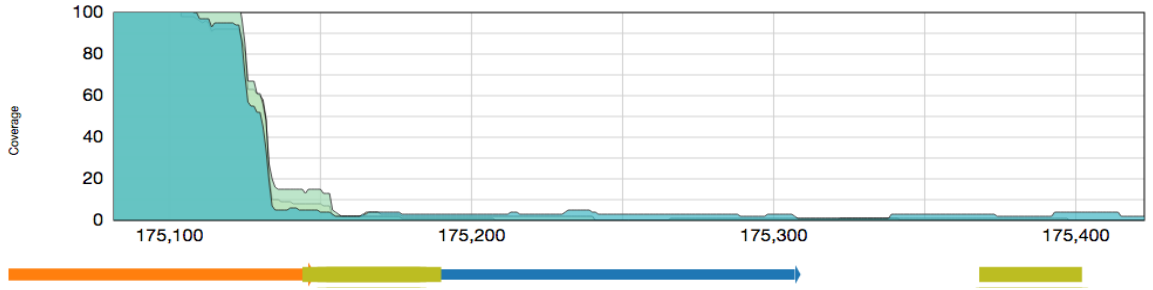
The bdh locus encodes two homologous butanol dehydrogenase(BDH) enzymes in a 3kb region on the main chromosome. Early studies of butanol dehydrogenases in *Clostridia* located a number of NADH-dependent and NADPH-dependent butanol and alcohol dehydrogenases responsible for butanoate metabolism[52;53;162;163],



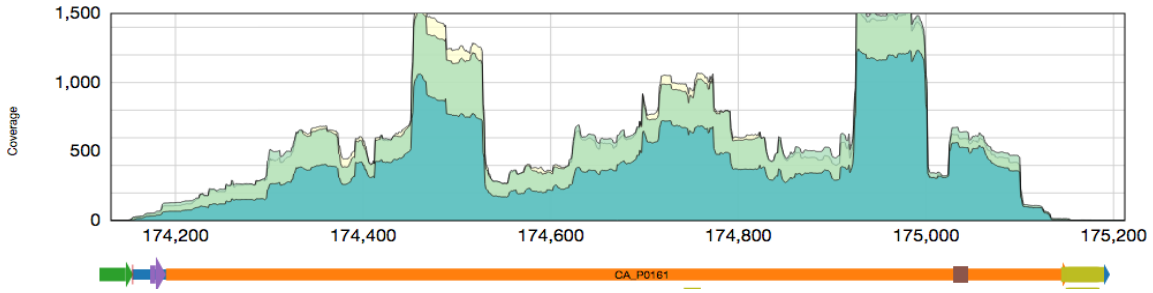
(a) solR Transcription Initiation Region



(b) solR Transcript



(c) solR Transcription Termination Region



(d) Curated solR Transcript

Figure 5.12: SolR Locus
a) The assembled transcription start site for solR agrees with previous findings. b) The assembled solR transcript has an extended 3' UTR and subreffig:5.12c) insufficient coverage for a transcript past the first terminator. d) The curated solR transcript agrees with previously published findings[157].

specifically the reduction of butyryl and acetyl groups into the solvents butanol and ethanol. One such locus in *C. acetobutylicum* produces two isozymes with different physiological roles. These isozymes like have distinct regulation and physiological roles from the other alcohol dehydrogenases found in this organism. The bdh locus proteins were described by several authors, reporting different activities and specificities for each enzyme[162;163]. After characterizing the enzymes in this locus, the region was cloned and two homologous isozymes were found. The two transcripts originating from these isozymes will demonstrate the precision and accuracy of this technique when compared to primer-extension analyses but also the need for assembly curation that reflects the motifs and coverage pattern of the region. We begin by discussing the first of these, bdhA.

5.3.2.1 BdhA

BdhA is an NADH-dependent butanol dehydrogenase that acts on both butyryl and acetyl groups. Studies suggest that this enzyme has fairly comparable activities with both substrates, with slightly higher activities for butyryl groups[162]. This enzyme was observed to have higher activities at low pH, indicative of its physiological role in the conversion of butyric acid to butanol. The entire locus was sequenced, producing ORFs that exactly matched the bdhA and bdhB isozymes[53]. Northern analysis determined that these genes are transcribed separately and not as an operon. BdhA was found to have a 1.3kb transcript and the transcription start site was mapped through primer-extension to base 3,465,240 on the main chromosome of *C. acetobutylicum*[53].

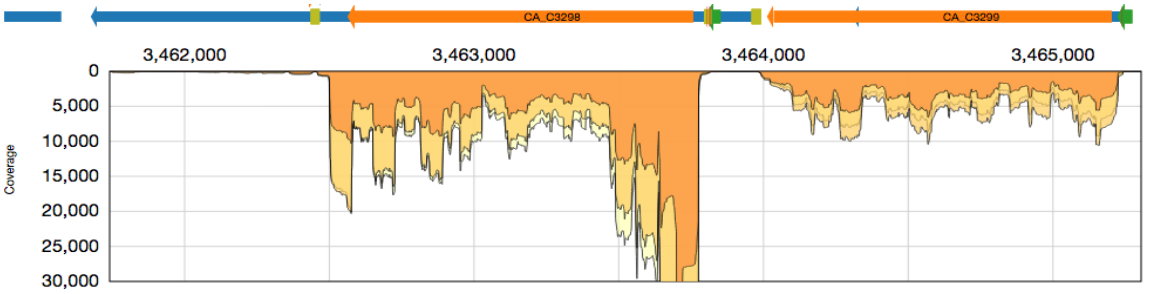
Here we find the transcription start site of bdhA one base upstream at 3,465,241(5.13b). The results of transcription start site identification agree well with the upstream

Sigma-factor A promoter and the aforementioned start site. The uncurated assembly produces a transcription stop site at base 3,464,329, before the stop codon of BdhA(5.13c). The pattern of coverage clearly reflects the Rho-independent terminator nearby. The misassembly could be due to low complexity in the reads mapping to this portion of bdhB, possibly resulting from homology with bdhB. This type of misassembly will be referred to from here on as a "truncation" misassembly. The final transcript (5.13d) reflects the assembly, coverage pattern, and motifs in this locus, agreeing with the transcription start site and a transcript length of 1,282bp. In this example, the fairly obvious coverage pattern was not reproduced by the assembly, demonstrating the need for some simple curation by integrating knowledge of previous experimental data and genome-wide predictions of promoter and terminator motifs with the coverage pattern and assembly. Next, the paralogous gene BdhB produces a similarly sized transcript.

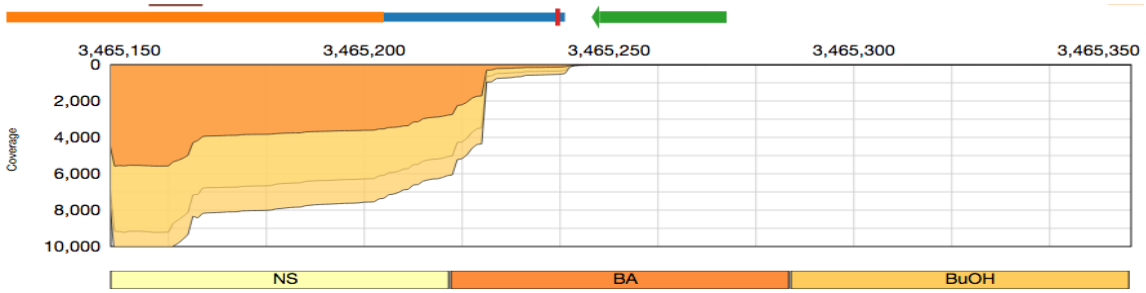
5.3.2.2 BdhB

The next gene is bdhB, another NADH-dependent butanol dehydrogenase, with a slightly longer transcript size of 1.35kb. It was reported that this enzyme has 46-fold higher activity with butyryl groups than acetyl groups[53;162]. BdhB was sequenced an analyzed along with bdhA, where it was discovered that bdhB had at least two transcription start sites, independent from BdhA. The most dominant transcription start site was very close to a secondary band at approximately 3,463,816 and 3,463,811, respectively[53]. I will refer to these two distal bands collectively as the primary transcription start site for bdhB. A third band was located slightly farther upstream at 3,463,803[53].

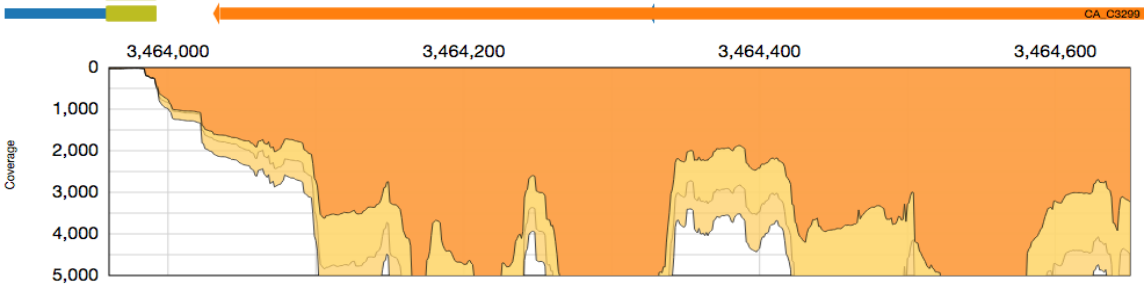
The coverage pattern for this region shows at least 3 increases in coverage at 3,463,802, 3,463,813, and 3,463,843 (5.14a). The first (proximal) site has a promoter



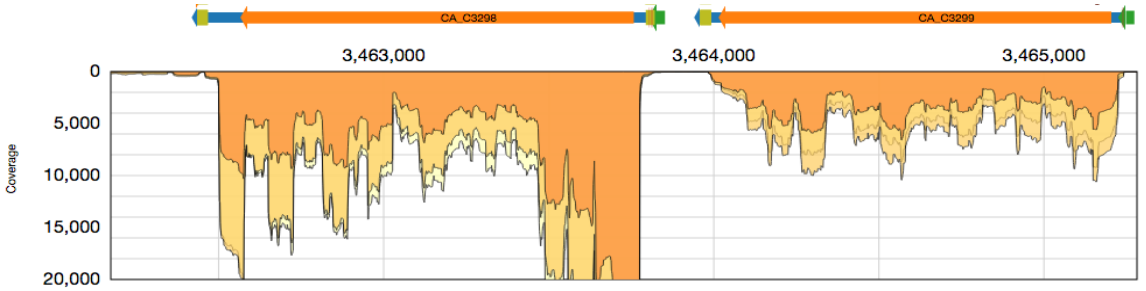
(a) bdh Locus



(b) bdhA Transcription Initiation Region



(c) bdhA Transcription Termination Region



(d) Curated bdhA Transcript

Figure 5.13: Bdh Locus

a) The bdh locus displays an obvious coverage pattern for two monocistronic transcripts. b) The bdhA transcript displays a sharp increase in coverage near the transcriptional start site. This data agrees to a good extent with primer extension studies for this gene. c) The raw assembly has failed to recapitulate the transcription termination region, likely due to low complexity coverage of the 3' region of this transcript. d) The curated transcript reflects experimental characterization of this transcript [53].

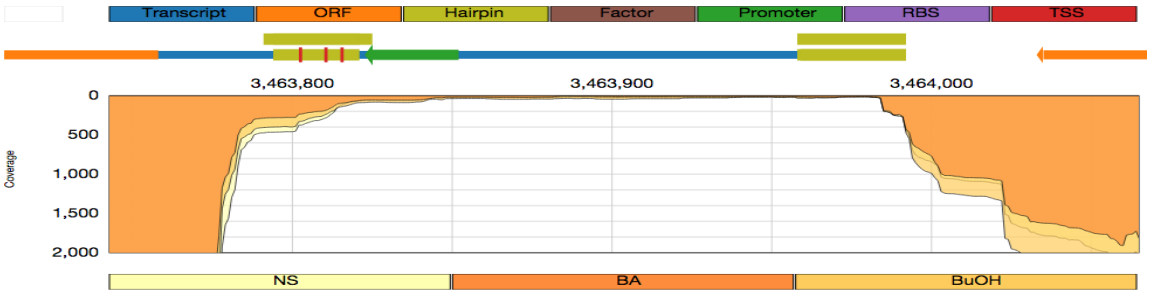
-35 box				
Motif	Start	End	Sequence	p-value
1	3463831	3463836	TAGGTT	3.5×10^{-2}
2	3463847	3463852	TTGTAA	9.4×10^{-3}
3	3463870	3463875	TGGATA	2.6×10^{-2}

-10 Box				
Motif	Start	End	Sequence	p-value
1	3463816	3463821	TATAAT	4.3×10^{-4}
2	3463820	3463825	TATATA	1.6×10^{-3}
3	3463830	3463835	TAAAAT	4.2×10^{-3}
4	3463852	3463857	TATTAT	4.2×10^{-3}

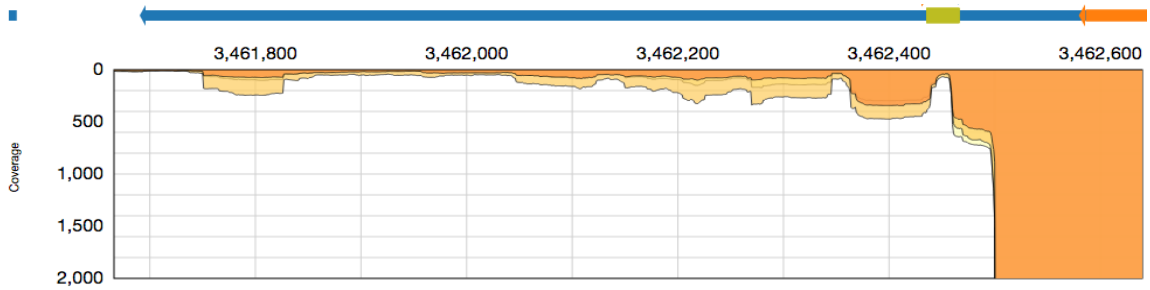
Table 5.2: BdhB Sigma-factor A boxes

motif upstream, corresponding to the 1^{st} -10 and -35 boxes in 5.2 and the tertiary transcription start site from above[52]. The second site has a significant increase in coverage between residue 3,463,817 and 3,463,810, which match well with the primary transcription start site described above[53]. The primary transcription start site matches the 2^{nd} -35 box and the 2^{nd} or 3^{rd} -10 box motifs[53]. It is clear that the strongest start site is the primary transcription start site previously described[53], although the multiple bands observed in their analysis could indeed be explained by the additional matches to consensus Sigma-factor A motifs. Additionally, an observable but insignificant increase is correlated with a final promoter motif. It is clear that the raw data match previous results to a good extent but curation was required to correct the transcription start site(5.14a) and stop site (5.14b). The final transcript is shown in 5.13c with a final length of 1,367bp or 1,381bp, in close agreement with the published length of 1.35kb.

In this example, the bdhA and bdhB transcripts were very close to capturing the true boundaries of these genes. In the case of bdhA the TSS was both precise



(a) bdhB Transcription Initiation Region



(b) bdhB Transcription Termination Region

Figure 5.14: Bdh Locus and Transcription Start Sites

a) The bdhB transcript has several promoter motifs and matching increases of coverage. These increases agree with previous experimental results[53], 2 which were dismissed after not identifying the appropriate promoters. Unfortunately, the transcription start site was incorrect in the raw assembly. b) A Rho-independent terminator is found at the end of the bdhB transcript although residual coverage triggered misassembly.

and accurate, but the assembled transcript did not match the coverage pattern, ORF, and terminator annotations, the first example of a truncation misassembly. It is possible that this and similar misassemblies could be due to homology between paralogues, such as in the C-terminus of bdhA and bdhB. In the case of bdhB, the start and stop sites did not agree with patterns in coverage (extension misassembly) and were simple to correct. The next example is another positive example, this time of a stress-response operon containing the heat-shock proteins GroES and GroEL.

5.3.3 GroES/EL Locus

5.3.3.1 GroES/EL Operon

The GroES and GroEL proteins are evolutionarily conserved heat-shock responsive chaperonins. These proteins are found throughout the tree of life, including *C. acetobutylicum*, where they are an integral to the solvent stress response[55;72;73] and the class I heat-shock response[58;72;73]. Expression levels are both strong and constitutive for this region, with a dramatic and transient increase with solvent or heat stress[72;73]. In a molecular study, GroES and GroEL were found to be produced from a bicistronic operon with a transcript size of 2,150bp[55]. In addition, a Sigma-factor A promoter was identified for an experimentally determined transcription start site. Two bands were observed during the primer-extension assay[55]. The proximal band was dismissed as an artifact although the bands persisted under all conditions examined[55]. This region is known to contain a CIRCE motif upstream[60], overlapping both of these start sites (**??**)[73].

The transcription start site determined in this work is located at 2,829,142, a mere two bases away from a previously reported start site[55] (5.15a). Interestingly, the second transcription start site is located at a sharp increase in expression. The

distance of this start site from the promoter would suggest either sequencing difficulties or post-transcriptional processing. The transcription stop site is located near a Rho-independent terminator (5.15b). The transcript size determined by the raw assembly is 2,131bp in agreement with the 2.2kb band and the 2,150bp calculated distance between the transcription start-site and the Rho-independent terminator[60]. In this example, the uncurated assembly produced a flawless recapitulation of the coordinates and size of the groES/EL operon (5.15b). Having established the coordinates for this transcript in agreement with previous findings, the regulation of this operon should be briefly discussed.

5.3.3.2 GroES/EL Regulation

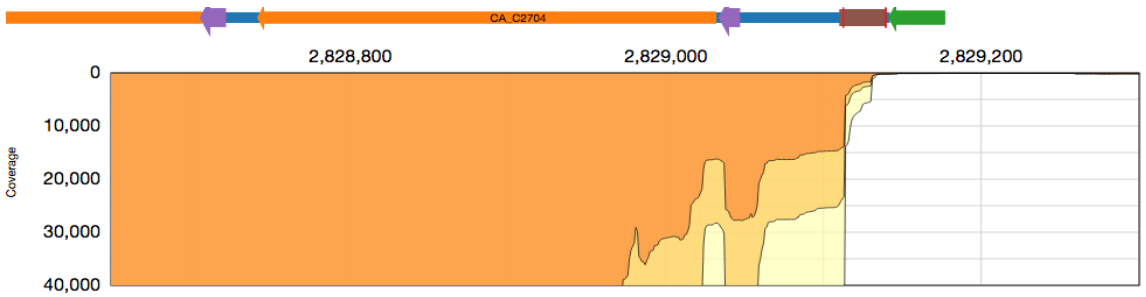
The groES/EL operon is stress responsive, although its expression under stress appears lower in 5.15c. For this reason, it is worth discussing the regulation of this region and why these results are consistent with knowledge of this area. The CIRCE motif upstream of groES is regulated by HrcA, a heat responsive repressor[58;60;74]. In response to heat-shock, the groES/EL operon is derepressed, revealing a Sigma factor A promoter and resulting in transcription (5.15a). The response to heat shock is fairly acute, increasing for 2-3 minutes and returning to standard levels after an additional 10 minutes[60]. Here a general decreased expression is observed in response to solvent stress for two reasons. First, the 6S small RNA is a stress and growth-stage responsive regulator that globally reduces transcription from Sigma factor A promoters[44;164]. Additionally, the time scale assessed here does not include a 3 minute time-point to identify an acute stress response. The stress response observed is the global downregulation of Sigma factor A promoters and the transition of the transcriptome towards one designed to tolerate stress and Sigma factor A dependent promoters are downregulated throughout this dataset as a result of the 6S small

RNA. For this reason, we defer the test of differential expression for these time points for the next chapter. In the case of the GroES/EL operon, we observe precise and accurate estimates of transcript boundaries from the uncurated assembly. The next example is the regulator responsible for the acute stress response of these heat shock proteins, HrcA.

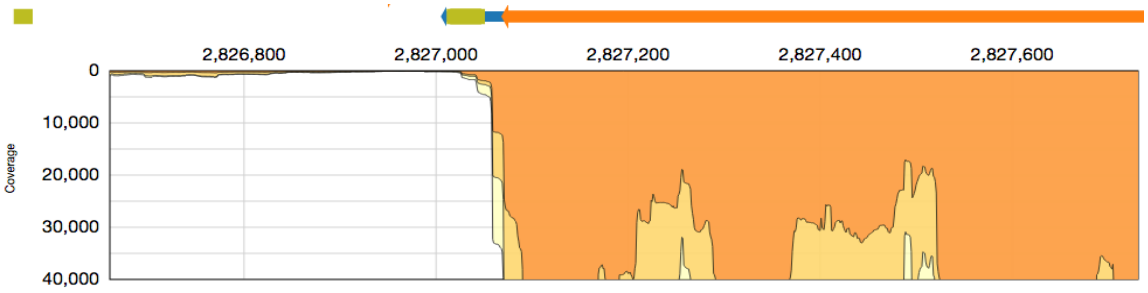
5.3.4 HrcA and DnaK/J Locus

5.3.4.1 DnaK Locus Overview

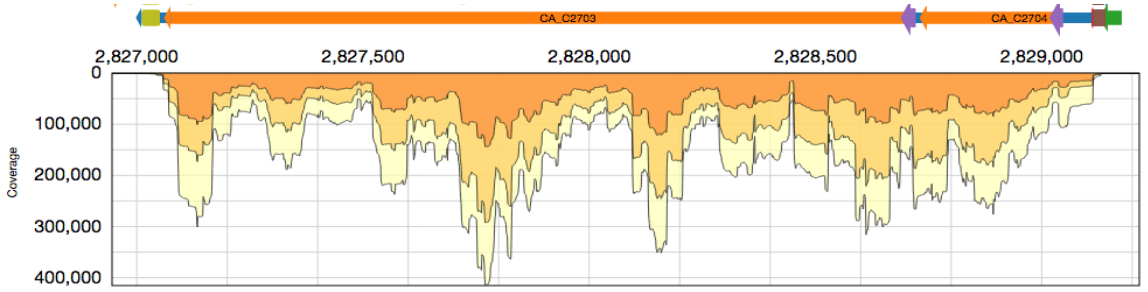
This rather complex locus encodes the class I repressor HrcA, DnaK and DnaJ, another set of evolutionarily conserved heat-shock proteins. DnaK was discovered to be solvent-stress responsive in *C. acetobutylicum*[72;73]. Solvent stress and heat shock increase protein denaturation, requiring molecular chaperones such as GroES/EL and DnaK/J to increase proper protein folding in these conditions. The *C. acetobutylicum* DnaK protein was purified as a stress responsive 74kDa protein[73]. Using a restriction fragment, the DnaK locus was then cloned and sequenced[56], revealing a grouping of four ORFs, similar in arrangement to *B. subtilis*[60]. An inverted repeat, now known to be the HrcA-binding CIRCE motif, was found upstream of the hrcA gene[56]. Upon Northern analysis of this region, three different transcripts were identified as originating from this locus, of 2.6, 3.8, and 5kb lengths. Additionally, a 1.5kb band was observed with a dnaK specific probe[56]. The first transcript (2.6kb) could be seen with a dnaK-specific probe and is thought to contain the genes grpE and dnaK. The second transcript (3.6kb) was observed for both dnaK and hrcA-specific probes and is thought to contain the same genes plus hrcA. Similarly, the 5kb transcript was observed for both probes and is thought to contain the whole operon. The final transcript(1.4kb) and additional small bands were dismissed as specific degradation products. It has been noted that this operon has interesting



(a) GroES/EL Transcription Initiation Region



(b) GroES/EL Transcription Termination Region



(c) GroES/EL Locus

Figure 5.15: GroES/EL Locus

a) groES and groEL form an operon that is responsive to heat-shock through a HrcA-mediated derepression mechanism. The transcription start sites are in agreement with previous findings with the addition of an interesting peak inbetween these two. b) The transcription termination region supports previous reports of a 2.2kb transcript terminating at a Rho-independent terminator following the GroEL ORF.

post-transcriptional regulation in *B. subtilis*, producing multiple transcripts from a heptacistronic operon[165]. To analyze this complex operon, we will proceed through 4 proposed regulatory sites. The first is the promoter region of hrcA. The second site is a transcription start site upstream of grpE. The third is the location of an internal CIRCE motif, terminator, and an additional transcription start site ahead of dnaJ. The final site is located at the end of the whole operon.

5.3.4.2 HrcA Promoter

The hrcA promoter was described during the sequencing of the dnaK/J locus[56]. This promoter produces the full transcript of 5kb and the smaller 3.6kb transcript terminating between dnaK and dnaJ. Two transcription start sites have been described for this region[56]. Both of the two reported transcription start sites(S1 and S2) were located upstream of the CIRCE element, in contrast to groES/EL (5.16a). Additional bands present in the primer extension analysis were rejected similarly to the strand bands in the groES/EL operon. An excellent Sigma factor A motif was located for the S1 site (P1,5.3) but only an insignificant motif ($p > 0.05$, TTTATG(17)AAAGAT) motif was found for the weak band of the S2 site. An alternative promoter (P2, 5.3) seems to be too close to the S2 transcription start site. If the observed bands represent true transcription start sites for this operon they are transcribed from close and overlapping promoter motifs.

Here we don't see any direct increase in transcription for the S1 site, most likely due sequencing difficulty near the CIRCE motif. Upon TEX enrichment, no increase or decrease in coverage can be observed near these sites(data not shown), suggesting that post-transcriptional processing is not responsible for these transcription start sites. The increases in transcription observed here are relatively minor

compared to the transcription of the entire HrcA operon. The P1 and P2 motifs seem to be sufficient for transcription initiation, although the coverage pattern shows evidence of complication of sequencing. In several studies[55;56] transcription start sites have been discarded due to local RNA secondary structure. It is reasonable that reverse transcription in this area may be complicated by the -10kcal/mol hairpin and CIRCE motif near the 5' end of the transcript.

Several authors have noted that full hrcA/dnaK operon transcripts are present at lower abundance[56;165], resulting in lower coverage and an indistinct transcription start site when considering coverage alone. In 5.16a, it is clear that the uncurated assembly estimated the transcription start site more accurately than a coverage-only approach would provide. As we have seen, in some cases the de Bruijn graph assembly requires curation to most accurately reflect local motifs. On the other hand, this method produces a reasonable estimate of the start site. After minor curation, the transcription start site agrees with the previously described[164] S2 and S1 (5.16b). Having established the transcription start site for the entire hrcA operon and previous transcriptional start sites are not the result of post-transcriptional processing, we move on to the higher abundance transcripts produced by this region, beginning with the grpE operon.

5.3.4.3 GrpE Promoter

The GrpE protein is an essential nucleotide exchange factor for DnaK. This protein was discovered upstream of dnaK after cloning and sequencing of the dnaK locus[56]. It was postulated that a second transcription start site upstream of grpE would explain the smaller 2.6kb transcript and a transcription start site was determined[56]. A transcription start site exists ahead of the grpE gene in *B. subtilis* as well[165]. The proposed promoter ahead of grpE does not match the Sigma factor A

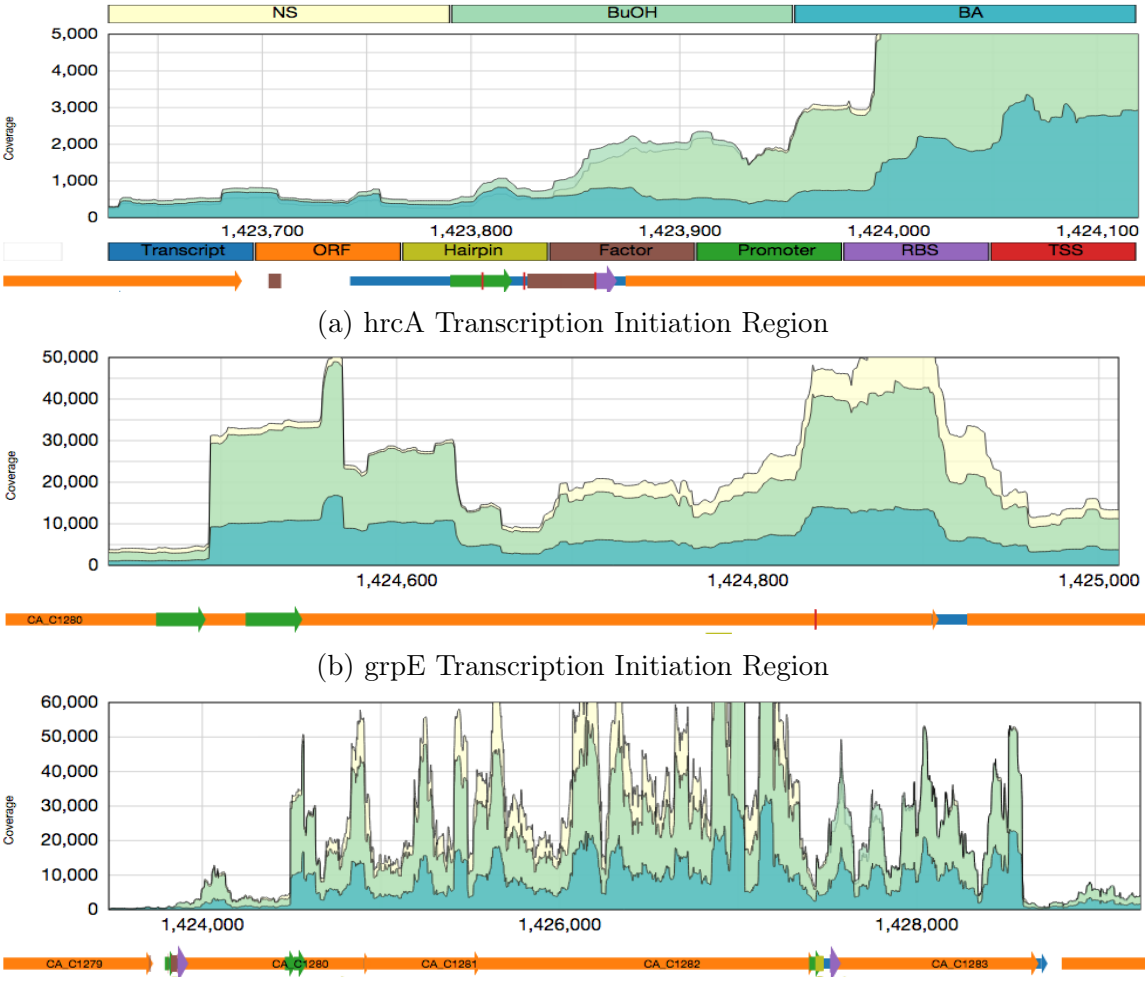
consensus ($p > 0.05$) and no alternative motifs were found. However, a substantial increase in coverage can be found further upstream from this site.

The first two sites are specifically enriched in the TEX treated library, suggesting a transcription start site. Promoter motifs of higher quality are found upstream of two major peaks (5.16b). Looking at the entire operon (5.16c) it is clear that this is a substantial transcription start site. The increase in coverage here is acute in contrast with the previous site in the conditions of this study. Additionally, it has been proposed that the hrcA operon is generated from post-transcriptional processing and differing transcript abundances are due to differing decay rates[165]. If such a processing mechanism were present in *C. acetobutylicum*, the coverage at the grpE transcription start site would be differential with respect to the 5'-phosphate exonuclease (TEX) treatment. In 5.16b, the coverage is not differential at this or other locations in this transcript, demonstrating that the 2.6kb transcripts is most likely a primary transcript originating from the Sigma factor A promoters P3 or P4.

In summary, a novel transcription start site and promoter motif were located upstream of grpE. No matching promoter motif or coverage pattern was observed for a previously documented transcription start site under these conditions. Additionally, coverage at the novel transcription start site was enriched with TEX treatment, suggesting that the 2.6kb transcripts are primary transcripts and not products of post-transcriptional processing. CtsR and CIRCE motifs were not found near this transcription start site. Next, the dnaK/J intergenic region is explored, which contains a terminator responsible for read-through transcription of the entire 5kb operon.

5.3.4.4 DnaK/J Intergenic Region

The dnaK/dnaJ intergenic region is known to contain a Rho-independent terminator[56;165], thought to be responsible for the 3.8kb and 2.6kb transcripts. The



(a) hrcA Transcription Initiation Region

(b) grpE Transcription Initiation Region

(c) Curated hrcA Locus

Figure 5.16: HrcA Locus and Promoter Regions

a) HrcA leads the 5kb tetracistronic operon. The regulation of this operon consists of Sigma factor A dependent promoters, a CstR motif, and a CIRCE motif. b) A secondary transcription start site in upstream of grpE is responsible for the 2.6kb and 3.8kb transcripts. A large increase in coverage at a novel transcription start site is not TEX responsive. c) The curated hrcA operon has two transcription start sites and one Rho-independent terminator which explain the 5kb, 3.8kb, and 2.6kb transcipts reported for this area.

-35 box				
Motif	Start	End	Sequence	p-value
P1	1423790	1423795	TTGACA	2.9×10^{-4}
P2	1423774	1423779	ATGAAA	5.3×10^{-2}
P3	1424463	1424468	TTGAGG	1.6×10^{-2}
P4	1424514	1424519	TTGATT	6.2×10^{-3}
P5	1427399	1427804	TTGAAA	2.1×10^{-3}

-10 Box				
Motif	Start	End	Sequence	p-value
P1	1423812	1423817	TATTTT	2.3×10^{-2}
P2	1423800	1423805	TAATGT	1.8×10^{-2}
P3.1	1424476	1424481	TAATAT	9.9×10^{-3}
P3.2	1424482	1424487	TAAAAA	3.2×10^{-2}
P4	1424537	1424542	TATGAT	1.9×10^{-3}
P5	1427430	1427435	TATAGT	2.5×10^{-3}

Table 5.3: HrcA Operon Sigma-factor A boxes

*Delta*G of this terminator is estimated to be -13.2kcal/mol. In 5.17a, decreased coverage is observed near this terminator, upstream of the dnaJ gene. This terminator does not contain a CIRCE element, in contrast to the inverted repeat at the hrcA promoter. A strong promoter motif (P5,5.3) is observed very close to a sharp increase in coverage at the 5' end of the repeat. The previously reported band is found on the 3' end of the repeat, which is explained by the strong terminator motif, similar to bands near the start of groES/EL and hrcA. To my knowledge, no dnaJ-specific Northern blots have been produced to identify a 1.2kb monocistronic transcript and thus, whether the promoter and observed increase in coverage reflect a true transcription start site remains unknown. Possible alternative explanations include post-transcriptional processing or thermodynamic challenges for the reverse transcriptase[56]. However, the coverage pattern does not show a response to the TEX

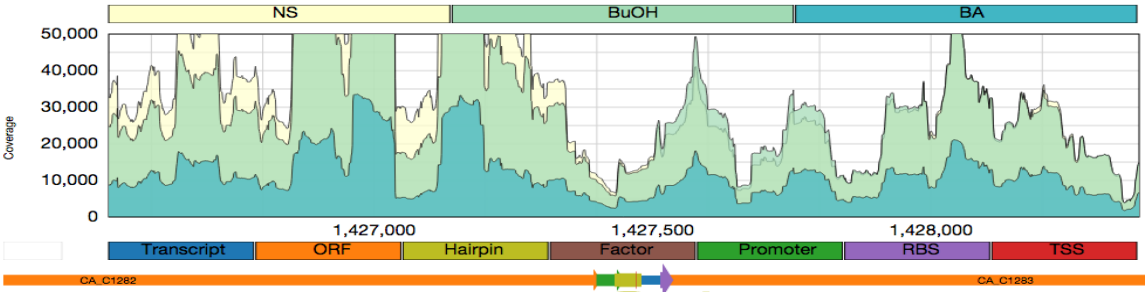
treatment (5.17a) and the former seems unlikely here. In the dnaK/J intergenic region, a decrease in coverage is observed near a strong Rho-independent terminator. Subsequently, increased coverage is observed near a strong promoter motif that might explain previous primer-extension results[56].

5.3.4.5 HrcA Transcript Termination

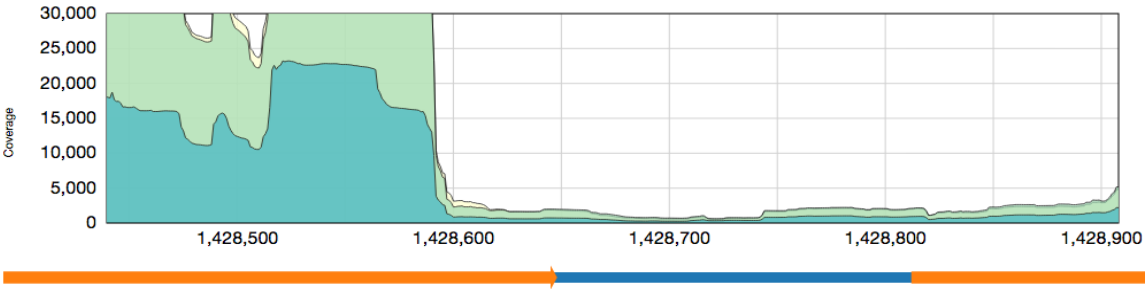
The full operon is produced in a 5kb transcript which terminates after the dnaJ gene. In 5.17b, a dramatic decrease in coverage is observed at the end of the dnaJ gene, 50 bases before the stop codon. Four terminator prediction software did not produce results for the transcription termination region shown here. This location should contain a non-intrinsic termination signal to explain the dramatic decrease in coverage. The residual signal from the nearby ribosomal methyltransferase(CA_C1284) matches well with evidence of a longer 8kb transcript in *B. subtilis*[165]. Given the dramatic decrease in coverage observed here, a reasonable transcription stop site for the 5kb operon can be assumed to follow the dnaJ gene(5.16c). Having discussed the regulatory regions of the hrcA operon, the final task is to summarize the transcript lengths, their regulation, and identify missing regulatory elements.

5.3.4.6 HrcA/GrpE Transcript Lengths

In the hrcA locus, two transcription initiation regions and two termination regions are responsible for 3 previously reported transcript sizes, 2.6, 3.8, and 5kb, respectively. The longest transcript observed here is 4,838bp from the hrcA transcription initiation region to the termination region downstream from dnaJ. The second largest observable transcript begins at one of the two start sites from the grpE transcription initiation region and ends at dnaJ termination region, leading



(a) DnaK/DnaJ Intergenic Region



(b) HrcA Transcription Termination Region.

Figure 5.17: HrcA Locus Transcription Termination Regions
a) The dnaK/dnaJ intergenic region consists of a Rho-independent terminator for the 2.6kb and 3.8kb transcripts. The coverage this region rebounds after a promoter motif and is not depleted with TEX treatment. b) Near the end of the dnaJ gene, a dramatic decrease in coverage signals the end of the 5kb hrcA operon. No terminators can be found in this region, suggesting a non-intrinsic termination signal.

to transcripts between 3,778 to 4,125 bases, respectively. This region is the most abundant portion of this operon and may contain at least two smaller transcripts. The first of these species would begin at the grpE transcription initiation region and terminate in the dnaK/J intergenic region. This type of transcript would range in size between 2,642bp and 2,989bp, respectively. The fourth and final transcript could originate from the dnaK/J intergenic region and terminate at the end of dnaJ, with a length of approximately 1.2kb. In the model organism *B. subtilis*, larger transcripts are produced from this region, up to 8kb in length[165]. It was observed that the transcripts and proteins of this region increase and decrease at different rates[165]. Determining all the potential transcription start sites inside a region of continuous transcription is an ongoing challenge for understanding of the stress response. To conclude, the regulatory motifs and their locations in this region are summarized.

5.3.4.7 HrcA Locus Regulation

Both Sigma factor A or H dependent promoters could be observed in the transcription initiation regions, as detailed above (5.3). Sigma factor A promoters are under the control of the 6S small RNA, which has considerable expression in this study. Therefore, the activity of this locus must be considered with respect to global Sigma A promoter activity. Additionally, the hrcA transcription initiation region is under the direct control of both a CIRCE motif (3.2×10^{-13}) and a CtsR motif (1.8×10^{-7})[58]. The inverted repeat in the dnaK/J intergenic region does not match either motif and is therefore most likely a Rho-independent terminator. The dnaJ termination region (+/- 200bp from the dnaJ stop codon) does not contain a detectable Rho-independent terminator and thus terminates transcription in a non-intrinsic manner, as previously suggested[166]. The sequencing technique has produced excellent results for this region. The assembly produced a better estimate of the

transcription start site than would be expected from coverage alone and a novel transcription start site was identified. Additionally, exact coordinates of regulatory motifs match well with the observed transcription start sites. Finally, potential transcript sizes were discussed to the extent permissible with this technique and the complexity of this region. Next, the spo0A locus is considered an important gene in this organism that has not had the level of molecular investigation of the previous examples.

5.3.5 Spo0A Locus

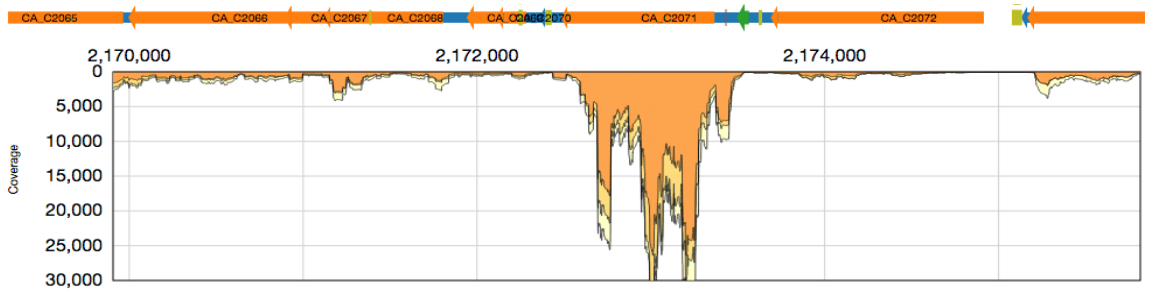
5.3.5.1 Spo0A

Spo0A is the master regulator of sporulation and stationary phase phenomena. This protein transduces growth-limiting and stressful signals into sporulation behavior in a number of firmicutes. In previous studies, Spo0A was shown to be translated from a 0.9kb transcript in *C. acetobutylicum*[167]. Additionally, a Sigma factor A and Sigma factor H motif were identified upstream of spo0A, but sadly neither the motifs nor the coordinates were provided. A single Sigma factor A motif (5.4) was identified near a substantial increase in transcriptional activity (5.18a). A single Spo0A box was reported upstream of spo0A[167] and can be seen in 5.18a. A single terminator motif is located 130 basepairs downstream of the spo0A stop codon. The uncurated assembly did not identify single start or stop sites for this gene, fusing spo0A with signal on either side of the gene. After curation with these information combined, we observe a transcript of 1,147bp (5.18b), longer than the 0.9kb transcript detected by Northern blot in a previous study of this locus[167]. This represents the first documentation of the Spo0A transcript boundaries in *C. acetobutylicum*.

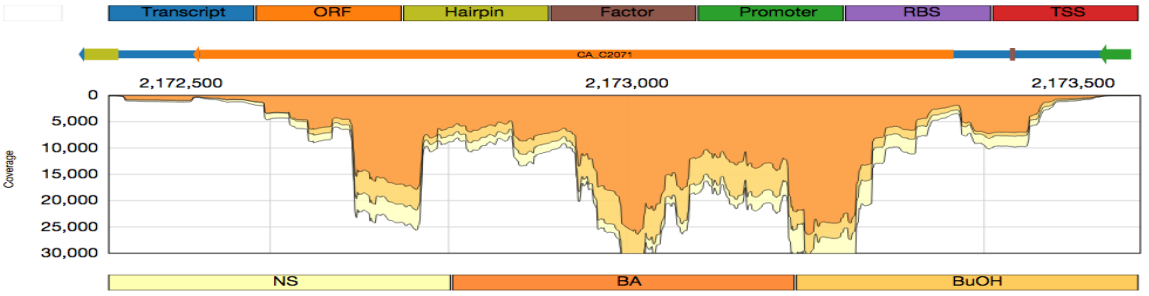
Sigma A				
Motif	Start	End	Sequence	p-value
-35	2173565	2173560	TTGATT	6.2×10^{-3}
-10	2173537	2173542	TAAAAT	1.5×10^{-3}

Spo0A Box				
Motif	Start	End	Sequence	p-value
1	2173430	2173436	TGTCGAA	1.9×10^{-4}

Table 5.4: Spo0A Regulatory Motifs



(a) spo0A Locus



(b) Curated spo0A Transcript

Figure 5.18: Spo0A Locus

??) The spo0A transcript is fused to signal from neighboring genes, although distinct promoter, terminator, and coverage signals are observed. **??**) After curation, this 1.1kb spo0A transcript reflects the appropriate genomic signals.

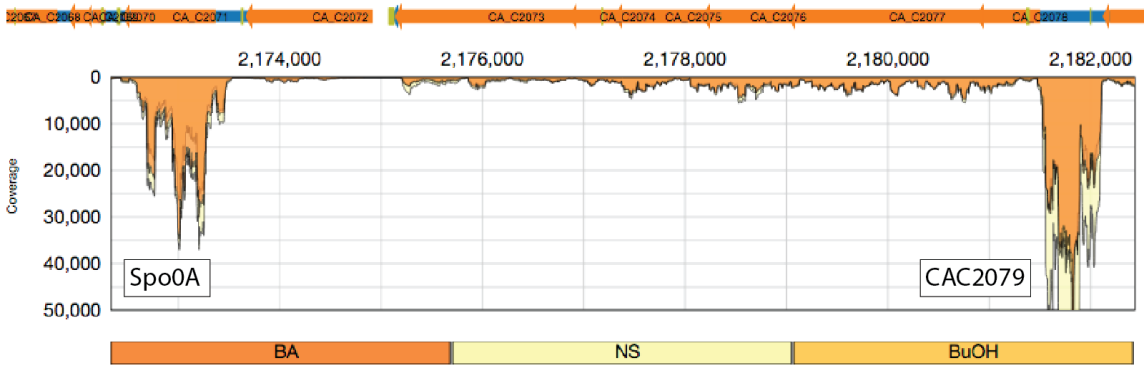


Figure 5.19: CA_C2079 Gene
Slightly downstream of the spo0A gene, a long operon follows a region of high expression on the Crick/minus strand. This coverage pattern is distinct from neighboring regions and surrounds a putative protein that is missing from the databases. This is the first experimental evidence for its expression and it shares homology to efflux transporters.

5.3.5.2 Missing from the Databases: CA_C2079

In a nearby location, the genes CAC2073-2078 are found in a tight grouping near a large peak of expression (5.19). This peak correpsonds to an uncharacterized protein CAC2079 that is present in UniProt but absent from NCBI and KEGG databases. Its substantial sequencing depth appears to be largely above 20k per base, independently transcribed from the surrounding regions. Bioinformatic analysis suggests that it shares sequence similarity to proteins in the *Clostridia*, *Bacilli*, *Baceteroidetes*, and *Halobacteria*. While there was no common catalytic or active domain unifying this group of homologs, a region of homology between them precedes a transmembrane motif. Further analysis via PSI-blast result suggests sequence similarity to mATE (Multidrug And Toxic-compound Extrusion) efflux family proteins. mATE family proteins use electrochemical gradients to export antibiotics and other toxic compounds. The data suggest that the expression of this protein is important to *C. acetobutylicum* and it will be interesting to examine the expression profile statistically. This final example illustrates the synergy of this transcriptomic dataset

with existing annotation and the benefit of curation for locating high priority novel transcripts.

In this section, several examples were presented to qualify the sequencing results. After the misassemblies were corrected, the transcript boundaries and sizes were in agreement with previous findings, suggesting precision and accuracy in this technique. In several cases, transcript boundaries required no curation at all. The *sol* locus illustrated the ability of the technique to even multiple transcription start sites with good accuracy. By integrating genomic and transcriptomic signals, the precision of the assembly was improved. After demonstrating the curation technique and validating results with prior studies, the curation technique and results are described in the next section.

5.4 Resolving Misassembly

These examples illustrated some common themes for misassemblies throughout the genome, notably the effect of background signal from high depth sequencing. Three types of misassemblies were described: extension, fusion, and truncation. In all cases these errors were resolved by solved by combining signals from sequencing depth and genome annotations. The assembly was fully curated for all pSol1 transcripts (the pSOL1 megaplasmid is 5% of the genome). This section explicitly states the rules and heuristics used for assembly curation. Also, the assembly was analyzed before and after curation to determine the effectiveness of this technique.

The curation process was used in the previous section to address the misassemblies, providing precision transcript boundary estimates. This technique used heuristics to correct transcript boundaries according to depth and annotations in

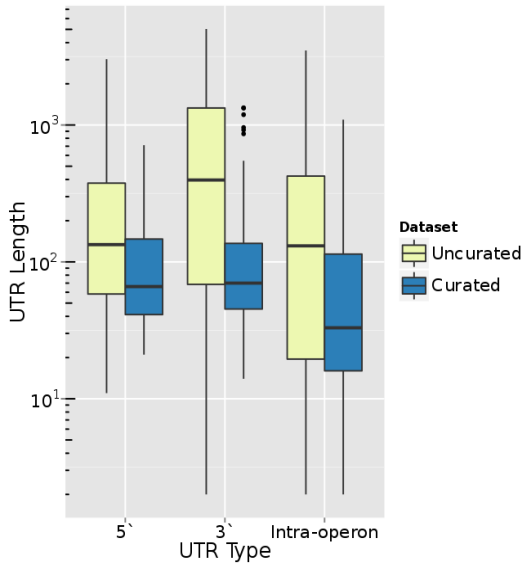
cases where interfering signals result in misassembly. Background signals were defined as extended (typically intergenic or antisense) regions of low depth that frequently conflicted with matching promoter/terminator annotations and large depth fold changes near transcript termini. Briefly, the heuristics are as follows:

1. Weak terminators (ΔG >-5 kcal/mol) were omitted.
2. Weak promoters and TFBSes were excluded from analysis when p $>1\times10^{-5}$, defined below for upstream and downstream motif matches (e.g. -35 and -10 elements of σ_A promoter).

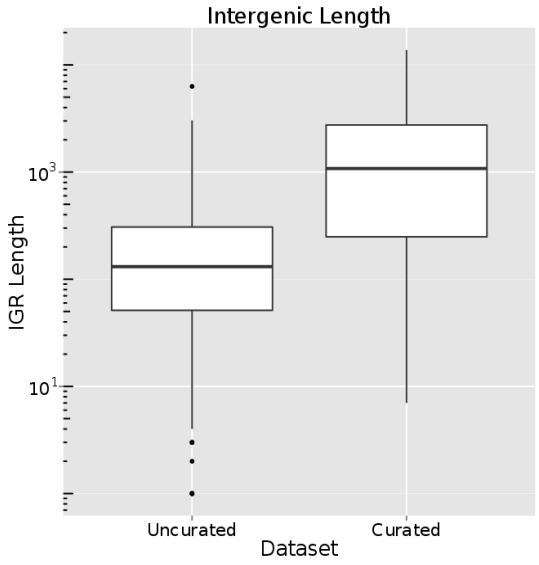
$$p_{promoter} = p_{upstream} \times p_{downstream}$$

3. Extended transcripts were corrected by shortening or spliting transcripts, such that the resulting transcript(s) captured depth patterns and annotations.
4. Fused transcripts were similarly addressed, with terminators as an important signal.
5. Truncated transcripts typically accompanied obvious trends in depth (e.g. BdhA) and were corrected by extension of the transcript to termini suggested by both depth and genome annotation.
6. Almost always, two or more signals (i.e. depth and terminator, etc.) in agreement were used to determine the true transcript boundaries. In edge cases, extended location specific differences in depth (>2 fold change) consistent across replicates were determined to be a transcript terminus.

These heuristics guided the curation process, addressing the errors described in previous sections, similarly to the treatment of the example transcripts. The most common types of misassemblies were the result of residual background signal assembled and mixed with true transcriptomic signal. The three types of errors were corrected during curation of the entire 192kb pSol1 megaplasmid, resulting in 111 transcripts spanning 192kb (5.5). In addition to improving the precision of transcript



(a) 5', 3', and Intraoperonic Untranslated Region Lengths



(b) Intragenic Region Lengths

Figure 5.20: UTR and IGR Lengths

a) UTR lengths were considerably improved by curation, with most less than 100bp, in agreement with *E. coli* averages.[156] Interestingly, a number of large 3' and intraoperonic regions remain after curation, suggesting either regulatory roles or the presence of unannotated proteins.

b) The size of intragenic regions (IGRs) increased upon curation after the drastic reduction in false-positive basepairs (5.5).

	Uncurated	Curated
Transcripts	181	110
Sequenced kb	347	190
Length Range	202-16kb	172-11kb
ORFs	155	157
Standard Transcripts	59	86
Standard kb	247	175
Novel Transcripts	122	24
Novel kb	100	15

Table 5.5: Final Assmelby Statistics and Curation Effect
This table shows final assembly statistics and the corrective power of the curation method. The number of misassembled baspairs and transcripts has been substantially reduced. Two additional ORFs/CDSes were included upon curation and a number of standard transcripts were split in half. Interestingly, about 20% of the assembled transcripts were novel, a good number of interesting candidates from a small portion of the total *C. acetobutylicum* genome.

boundary determination, the type I error for transcript discovery was reduced by removing a large number of false positive transcripts.

The transcript length distributions agreed with prokaryotic averages (5.22) after improved precision detailed above.[155] Specifically, the distribution of untranslated region lengths closely matched *E. coli* averages[156] (5.20a). In contrast to the decreased transcript and UTR lengths, intergenic region lengths increased upon curation (5.20b). The curated transcripts are evenly spaced along the pSol1 megaplasmid, in contrast to the uncurated assembly. Additionally, both the standard and novel transcripts have higher average per-base depth after curation, comparable to the coverage of the reference ORFs (5.21). These data show considerable agreement with *E. coli* averages and improvement over the uncurated assembly results.

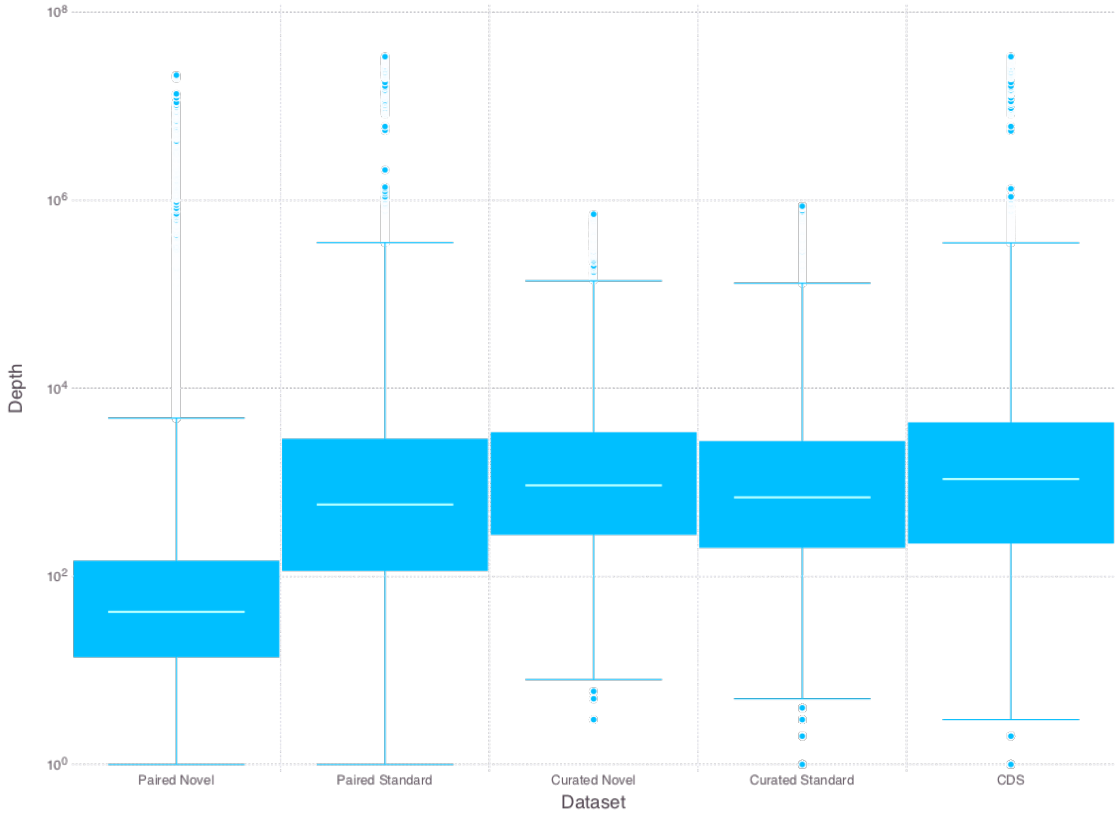


Figure 5.21: Cumulative Depth Distribution

After curation, the expression level as indicated by the per-base sequencing depth has increased substantially. While uncurated novel transcripts (far-left) had an order of magnitude lower average sequencing depth than found in reference ORFs (far-right, the 24 curated novel transcripts (center) had comparable sequencing depth. The depth of uncurated standard transcripts (middle left) were only slightly improved in terms of depth by curation (middle right). The best improvement in the novel transcripts was the increased precision of the final transcript boundaries.

5.5 First Strand-Specific Transcriptome Assembly for *Clostridia*

In this chapter, a strand-specific transcriptome assembly was conducted with the high-depth sequencing dataset obtained for *C. acetobutylicum* under various experimental conditions. The resulting boundaries were determined precisely due to laboratory and informatic quality controls and curation of the assembly. Significantly, these results represent the first strand-specific transcriptome assembly for the class *Clostridia*. Moreover, the innovative approach described here addresses misassemblies that result from the limitations of modern sequencing technology that are frequently neglected by similar studies.

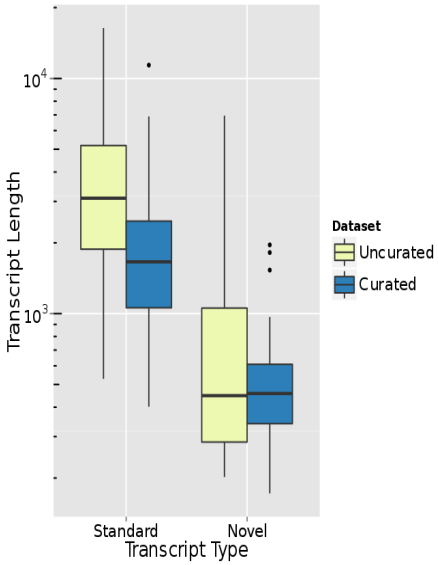


Figure 5.22: Transcript Lengths
The transcript lengths were improved after curation, centering the distribution on the standard transcripts on the *E. coli* average of 1.1kb[155].

The initial assembly from the subset of properly-paired reads had maximal size, expression, and inclusion of reference protein annotations compared to the assembly from the full dataset. However, several types of misassemblies were present in the dataset that were observed statistically and with specific examples. To correct these errors, a customized genome browser was developed for integrative analysis of complexity, sequencing depth, and genome annotations. Heuristics for curation of the assembly were described and exemplified with canonical *C. acetobutylicum* transcripts. In each case, the curation method produced transcript boundaries with precision comparable to previous targeted studies. The

curation method was applied to the entire pSol1 megaplasmid, fixing misassembled transcripts. The curation technique deeply improved the type I error rate for transcript discovery and boundary determination. The integrative analyis was resilient to background signals of the sequencing technique, resulting in a high-quality assembly for the pSol1 megaplasmid. Re-evaluation of the final assembly showed transcriptome feature lengths in agreement with prokaryotic averages. These data clearly suggest the efficacy of this technique and the precision of the results.

Chapter 6

CONCLUSIONS

Renewables research revolves around the development of a feedstock-flexible chassis organism that requires minimal engineering for biofuels production, such as *C. acetobutylicum*. The development of this organism requires a complete genome annotation consisting of ORFs, promoters, terminators, and transcript boundaries. The existing annotation of this microbe is largely the result of ORF predictions from antiquated gene models and is consequently incomplete. Genomic and molecular research will be more efficacious with an accurate genome annotation.

High-throughput transcriptomic methods such as RNA-seq are ideal to update this annotation, despite the well documented challenges related to this platform. Many of these challenges have not been addressed by the literature, leading to poor data utilization rates(Table 2.1). The absence of standards for bacterial transcriptome mapping studies provided the opportunity to develop an innovative technique to explicitly address false positive and false negative signal in sequencing datasets.

Several problematic issues, including rRNA and RNA degradation, were addressed by developing a laboratory workflow and quality control system prior to deep sequencing. The dataset was cleaned for errors, biases, and contaminants for proper quantification of the sensitivity. 450M properly-paired reads provided >9000 fold-coverage of the *C. acetobutylicum*, with a median per-base sequencing depth of 156x. This method even detected low-level background signals, an issue affecting

deep sequencing studies that leads to false positive errors, yet ignored by studies in the microbial community.

To identify and treat these issues, a fast and flexible genome browser was constructed. The genome browser visualized the background signals and misassemblies that were detectable in assembly statistics. Genome-wide promoter predictions revealed the prevalence of σ_A-promoter elements in the AT-rich *C. acetobutylicum* genome, a potential source for the background signals. An integrative analysis method was developed to correct the misassemblies where necessary by including sequencing depth, complexity, Rho-independent terminators, and promoter motifs in the annotation visualization and curation method.

Most examples required little to no curation and showed excellent precision and accuracy with respect to previous studies. Even challenging edge cases involving multiple transcription start sites had excellent signal to noise ratio and consequently simple corrections. A proof-of-principle curation of the pSOL1 megaplasmid produced ideal assembly statistics, including a median transcript size of 1.4kb consistent with the reported average transcript lengths in *E. coli*. A total of 86 reference-ORF containing transcripts and 24 novel transcripts were identified.

By explicitly addressing several issues related to false positive and false negative signals, a sequencing protocol and integrative analysis method was developed. This method lead to the first strand-specific transcriptome assembly in the genus *Clostridia*. The technique described by this work is applicable in any bacterial species where a genome sequence is available. While the unprecedented sequencing depth of this study lead to false positive results in the initial assembly, the integrative curation method provided both precision and accuracy in transcript boundary determination.

Chapter 7

FUTURE WORK

7.1 Annotation Completion and Differential Expression Analysis

This study provided *C. acetobutylicum* transcript boundaries for future molecular and genomic studies. The method used in the proof of principle curation of the pSOL1 megaplasmid should be extended to the entire *C. acetobutylicum* chromosome. It is reasonable to expect that the transcript and UTR sizes for the whole genome should be similarly improved. The transcript boundaries could improve expression estimates and differential expression analyses. Beyond comparing with previous microarray studies, novel transcripts discovered here could be discovered to be stress responsive. Such findings would be natural targets for future targeted or whole genome stress response analyses.

Specifically, this work paves the way for a refined differential expression study. Differential expression with RNA-seq relies on identifying statistically different counts of sequenced cDNAs between samples.[141] These counts can be acquired for each ORF individually, or for entire transcripts. The larger size of transcripts provides an increased sampling area for read counting. It is reasonable that this may provide a better representation of the expression of the RNA molecule and is preferable statistically to expression measurements from ORFs alone. Augmented with transcript boundaries and novel transcripts, this genome annotation could then be used for differential expression analysis of this and other datasets.

7.2 Annotation Cross-validation

A previous standard RNA-seq dataset from *C. acetobutylicum* could be used to further evaluate the transcripts identified here.[44] While I was a co-author on this paper, I served in a purely computational role and did not handle the RNA material prior to sequencing. Therefore, the dataset represents an independent measurement of the *C. acetobutylicum* transcriptome and could be used to cross-validate the transcripts. Detection of these transcripts by more than one dataset would provide additional verification and significance.

7.3 Further Misassembly, Background Noise Investigation

With a complete genome annotation, a particularly important comparison may be made: how does the sequencing depth in intergenic regions (background noise) compare with the identified transcribed regions? Is there a statistically significant difference? Furthermore, are the misassembled transcripts/bases all located in these intergenic regions or co-localized to specific regions of the genome? Were the improperly paired reads also co-localized? The presence of phage sequences in the *C. acetobutylicum* genome has been recently identified (unpublished results), which may have affected the alignment rates and/or co-localization of reads observed in this study. The answer to these questions can be obtained with a fully curated set of transcripts and computational investigation of the phage regions.

7.4 Regulatory Motif Investigation

The transcript boundaries provided by this work also facilitate regulatory motif identification. Besides the promoter motifs and transcription factor binding sites described here, new motifs could exist upstream of transcription start sites of clusters of co-regulated genes. Differential promoter usage can be investigated using the

genome browser and gene-specific techniques. Gene networks in *C. acetobutylicum* inevitably possess transcription activation systems and will be ready to be explored with a complete genome annotation.

In fact, differential expression and motif analyses can be combined in an interesting way. By coupling differentially expressed genes, their expression profiles, and clustering algorithms, patterns of co-expression and perhaps co-regulation may be identified. The performance of such an approach naturally depends on normalization approaches (to make profiles reasonably comparable) and clustering theory. While some approaches have been suggested for expression profiles specifically, the performance of any approach is ultimately determined by these two factors and thus all options should be explored.

With a set of likely promoters, comparison of these promoters to less-significant promoters located throughout the genome is desirable. There may be statistically detectable differences in promoter motifs that are desirable for refining the understanding of the *Clostridia* promoter.

7.5 Transcriptome Annotation

Additional research can be done with comparative genomics, including re-annotation of the genome and transcriptome for protein coding ORFs. With the example of the missing CAC2079 gene for example, there may be substantially transcribed and protein coding regions that require comparative analysis to identify metabolite exporters, two-component systems, and more. In fact, the RAST annotation system[106] can be used to annotate transcriptomes, with some clever scripting and knowledge of its features.

7.6 Additional Genome Browser Features and Deployment

The genome browser in its current state is most useful to those who are familiar with it and how it can be used most effectively. Additional features can be added to simplify this resource for more users. For example, the current user interface utilizes genomic coordinates for browsing as opposed to browsing by an individual gene (A.2). The addition of UI features and aesthetics for the browser might facilitate its adoption. Additional flexibility regarding the annotation uploading/editing process might improve its utility as well, although gtf remains a widely used format and is readily converted from BED format and others. Downloadable features such as conversion scripts and the complete annotation in gtf and/or genbank format may also facilitate its use.

Comparison of the PostgreSQL database with the less intensive MySQL format may reveal performance advantages that should be used. Additionally, MySQL is a widely adopted database format and as a result, may be simpler to deploy. Deployment of this resource could be accomplished locally at the University of Delaware, but additional sources could prove useful.

7.7 Complexity Index

A theme throughout this work was the benefit of integrating multiple perspectives of the dataset, specifically library complexity, sequencing depth, and bioinformatically-predicted motifs. However, the analysis of complexity was restricted to the assembly by the Trinity algorithm. Ultimately, library complexity is a function of the uniquely-sequenced cDNAs from a particular position in the genome. Trinity presents this information in a binary manner (i.e. active/assembled vs inactive/unassembled). Alternatively, a quantitative basepair-specific library complexity index would be an incredible useful metric to additionally understand the quality of transcription in a

region. This could be overlayed with sequencing coverage in a creative way to further illustrate the patterns that are useful for identifying truly transcribed regions.

7.8 Machine Learning Algorithm for Alternative to Transcriptome Assembly

Another theme in this work was the imperfection behind the transcriptome assembly paradigm. Realistically, sequencing datasets can be expected to have some basal levels of noise from DNA contamination or spurious transcription. Additionally, it seems that studies that do not dicuss this issue are not sufficiently deep. Therefore, in truly deep sequencing datasets, transcriptome assemblies can produce false positive errors (extra assembled bases and transcripts). Therefore, it is desirable to develop a bioinformatic procedure to automate the process described in this work for application as an alternative to transcriptome assembly.

Ideally, this procedure would be as independent as possible from genomic GC content and other characteristics. It would leverage a sequencing dataset from an organism with a reference genome. By converting aligned reads (BAM files) into coverage vectors, this information could be combined with promoter predictions from a related organism and terminator motif predictions to identify truly expressed regions. This technique could be used for cross-validation of this research and would be a simple alternative to the complex assembly curation process.

7.9 Small RNA Target Prediction

Finally, additional analyses should be done for RNA hybridization and small RNA target prediction. Many recently identified stress responsive small RNAs[44] have unknown targets and functions. Exploration of their roles has been prohibited by undefined UTR structures and thus the thermodynamics of interaction with

their partners. A complete genome annotation provides these information and an additional tractable problem for *C. acetobutylicum* researchers to explore.

The first assembly and annotation provided by this work presents a number of opportunities for additional work in *C. acetobutylicum*. Increasing temperatures, CO_2 levels, and energy prices provide ethical and economical incentives to explore renewables research in this organism. A small number of intrinsic conditions have explored in this study, including the exponential and transition stages of growth, and stress responses to metabolite stress. Additional conditions may display alternate gene sets, expanding the complexity of the transcriptome beyond what is described here. Novel transcripts identified in the megaplasmid and chromosome require annotation and molecular study. If enzymes are discovered, metabolic models should be updated. Differential expression experiments are an increasingly useful method for understanding transcriptomic dynamics and represent an opportunity to further explore this dataset. Such studies in *C. acetobutylicum* directly benefit from a wider and more complete genome annotation. This work facilitates future research in *C. acetobutylicum* through the genome browser, which can incorporate future annotations and display expression data in a fast, flexible, and data-dense manner.

BIBLIOGRAPHY

[1] TF Stocker, D Qin, GK Plattner, M Tignor, SK Allen, J Boschung, A Nauels, Y Xia, B Bex, and BM Midgley. *IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*, 2013.

[2] 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].

[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 *Escherichia coli* for 1-butanol production. *Metabolic engineering*, 10(6):305–311, 2008.

[5] 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.

[6] 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.

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

[8] John T Heap, Oliver J Pennington, Stephen T Cartman, Glen P Carter, and Nigel P Minton. The clostron: A universal gene knock-out system for the genus *Clostridium*. *Journal of Microbiological Methods*, 70(3):452–464, 2007.

[9] John T Heap, Sarah A Kuehne, Muhammad Ehsaan, Stephen T Cartman, Clare M Cooksley, Jamie C Scott, and Nigel P Minton. The clostron: Mutagenesis in *Clostridium* refined and streamlined. *Journal of microbiological methods*, 80(1):49–55, 2010.

[10] Mohab A Al-Hinai, Alan G Fast, and Eleftherios T Papoutsakis. Novel system for efficient isolation of *Clostridium* double-crossover allelic exchange mutants enabling markerless chromosomal gene deletions and dna integration. *Applied and environmental microbiology*, 78(22):8112–8121, 2012.

[11] Jennifer Au, Jungik Choi, Shawn W Jones, Keerthi P Venkataramanan, and Maciek R Antoniewicz. Parallel labeling experiments validate *Clostridium acetobutylicum* metabolic network model for ^{13} c metabolic flux analysis. *Metabolic engineering*, 26:23–33, 2014.

[12] Satyakam Dash, Thomas J Mueller, Keerthi P Venkataramanan, Eleftherios T Papoutsakis, Costas D Maranas, et al. Capturing the response of clostridium acetobutylicum to chemical stressors using a regulated genome-scale metabolic model. *Biotechnology for biofuels*, 7(1):144, 2014.

[13] 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.

[14] Alan G Fast and Eleftherios T Papoutsakis. Stoichiometric and energetic analyses of non-photosynthetic co_2-fixation pathways to support synthetic biology strategies for production of fuels and chemicals. *Current Opinion in Chemical Engineering*, 1(4):380–395, 2012.

[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] 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.

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

[18] Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S García-Sotelo, Alejandra López-Fuentes, et al. Regulondb version 7.0: transcriptional regulation of *Escherichia coli* k-12 integrated

within genetic sensory response units (gensor units). *Nucleic acids research*, 39(suppl 1):D98–D105, 2011.

[19] Tilman J Todt, Michiel Wels, Roger S Bongers, Roland S Siezen, Sacha AFT Van Hijum, and Michiel Kleerebezem. Genome-wide prediction and validation of sigma70 promoters in *Lactobacillus plantarum* wcfs1. *PloS one*, 7(9):e45097, 2012.

[20] Jan-Philip Schlüter, Jan Reinkensmeier, Melanie J Barnett, Claus Lang, Elizaveta Krol, Robert Giegerich, Sharon R Long, and Anke Becker. Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium *Sinorhizobium meliloti* 1021. *BMC genomics*, 14(1):156, 2013.

[21] Pieter Meysman, Julio Collado-Vides, Enrique Morett, Roberto Viola, Kristof Engelen, and Kris Laukens. Structural properties of prokaryotic promoter regions correlate with functional features. *PloS one*, 9(2):e88717, 2014.

[22] Jörk Nölling, Gary Breton, Marina V Omelchenko, Kira S Makarova, Qiandong Zeng, Rene Gibson, Hong Mei Lee, JoAnn Dubois, Dayong Qiu, Joseph Hitti, et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *Journal of bacteriology*, 183(16):4823–4838, 2001.

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

[24] 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.

[25] Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011.

[26] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[27] ENCODE Consortium. *Standards, Guidelines and Best Practices for RNA-Seq*, 2011.

[28] José A Robles, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using rna-sequencing. *BMC genomics*, 13(1):484, 2012.

[29] Yuwen Liu, Jie Zhou, and Kevin P White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

[30] Cynthia M Sharma, Steve Hoffmann, Fabien Darfeuille, Jérémy Reignier, Sven Findeiß, Alexandra Sittka, Sandrine Chabas, Kristin Reiche, Jörg Hackermüller, Richard Reinhardt, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 464(7286):250–255, 2010.

[31] Yi Wang, Xiangzhen Li, Yuejian Mao, and Hans P Blaschek. Single-nucleotide resolution analysis of the transcriptome structure of *Clostridium beijerinckii* ncimb 8052 using rna-seq. *BMC genomics*, 12(1):479, 2011.

[32] Shan Li, Xia Dong, and Zhengchang Su. Directional rna-seq reveals highly complex condition-dependent transcriptomes in *E. coli* k12 through accurate full-length transcripts assembling. *BMC genomics*, 14(1):520, 2013.

[33] Shaomei He, Omri Wurtzel, Kanwar Singh, Jeff L Froula, Suzan Yilmaz, Susannah G Tringe, Zhong Wang, Feng Chen, Erika A Lindquist, Rotem Sorek, et al. Validation of two ribosomal rna removal methods for microbial metatranscriptomics. *Nature methods*, 7(10):807–812, 2010.

[34] Hana Yi, Yong-Joon Cho, Sungho Won, Jong-Eun Lee, Hyung Jin Yu, Sujin Kim, Gary P Schroth, Shujun Luo, and Jongsik Chun. Duplex-specific nuclease efficiently removes rrna for prokaryotic rna-seq. *Nucleic acids research*, page gkr617, 2011.

[35] Clelia Peano, Alessandro Pietrelli, Clarissa Consolandi, Elio Rossi, Luca Petiti, Letizia Tagliabue, Gianluca De Bellis, and Paolo Landini. An efficient rrna removal method for rna sequencing in gc-rich bacteria. *Microb Inform Exp*, 3(1):1–11, 2013.

[36] 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.

[37] Jan Mitschke, Jens Georg, Ingeborg Scholz, Cynthia M Sharma, Dennis Dienst, Jens Bantscheff, Björn Voß, Claudia Steglich, Annegret Wilde, Jörg Vogel, et al. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis sp.* pcc6803. *Proceedings of the National Academy of Sciences*, 108(5):2124–2129, 2011.

[38] Shahriar Shafiee and Erkan Topal. When will fossil fuel reserves be diminished? *Energy Policy*, 37(1):181–189, 2009.

[39] L Bruce Railsback. Depth and nature of giant petroleum discoveries through time as an indicator of resource depletion. *Journal of Industrial Ecology*, 17(3):345–351, 2013.

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

[41] Benjamin G Harvey and Heather A Meylemans. The role of butanol in the development of sustainable fuel technologies. *Journal of Chemical Technology and Biotechnology*, 86(1):2–9, 2011.

[42] David T Jones and David R Woods. Acetone-butanol fermentation revisited. *Microbiological reviews*, 50(4):484, 1986.

[43] JR Martin, H Petitdemange, J Ballongue, and R Gay. Effects of acetic and butyric acids on solvents production by *Clostridium acetobutylicum*. *Biotechnology Letters*, 5(2):89–94, 1983.

[44] 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.

[45] Ioanna Pagani, Konstantinos Liolios, Jakob Jansson, I-Min A Chen, Tatyana Smirnova, Bahador Nosrat, Victor M Markowitz, and Nikos C Kyrpides. The genomes online database (gold) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(D1):D571–D579, 2012.

[46] Anna Azvolinsky. Sequencing the tree of life. *The Scientist*, 28(4), 2014.

[47] Ramesh V Nair, George N Bennett, and Eleftherios T Papoutsakis. Molecular characterization of an aldehyde/alcohol dehydrogenase gene from *Clostridium acetobutylicum* atcc 824. *Journal of bacteriology*, 176(3):871–885, 1994.

[48] Ralf J Fischer, Jan Helms, and P Dürre. Cloning, sequencing, and molecular analysis of the *sol* operon of *Clostridium acetobutylicum*, a chromosomal locus involved in solventogenesis. *Journal of bacteriology*, 175(21):6959–6969, 1993.

[49] Ulrike Gerischer and Peter Dürre. Cloning, sequencing, and molecular analysis of the acetoacetate decarboxylase gene region from *Clostridium acetobutylicum*. *Journal of bacteriology*, 172(12):6907–6918, 1990.

[50] Daniel J Petersen, Jeffrey W Cary, Jos Vanderleyden, and George N Bennett. Sequence and arrangement of genes encoding enzymes of the acetone-production pathway of *Clostridium acetobutylicum* atcc 824. *Gene*, 123(1):93–97, 1993.

[51] Lisa Fontaine, Isabelle Meynial-Salles, Laurence Girbal, Xinghong Yang, Christian Croux, and Philippe Soucaille. Molecular characterization and transcriptional analysis of *adhE2*, the gene encoding the nadh-dependent aldehyde/alcohol dehydrogenase responsible for butanol production in alcohologenic cultures of *Clostridium acetobutylicum* atcc 824. *Journal of bacteriology*, 184(3):821–830, 2002.

[52] DANIEL J Petersen, RICHARD W Welch, FREDERICK B Rudolph, and GEORGE N Bennett. Molecular cloning of an alcohol (butanol) dehydrogenase gene cluster from *Clostridium acetobutylicum* atcc 824. *Journal of bacteriology*, 173(5):1831–1834, 1991.

[53] KARL A Walter, GN Bennett, and ELEFTHERIOS T Papoutsakis. Molecular characterization of two *Clostridium acetobutylicum* atcc 824 butanol dehydrogenase isozyme genes. *Journal of bacteriology*, 174(22):7149–7158, 1992.

[54] John Wong, Catherine Sass, and George N Bennett. Sequence and arrangement of genes encoding sigma factors in *Clostridium acetobutylicum*> atcc 824. *Gene*, 153(1):89–92, 1995.

[55] FRANZ Narberhaus and HUBERT Bahl. Cloning, sequencing, and molecular analysis of the *groESL* operon of *Clostridium acetobutylicum*. *Journal of bacteriology*, 174(10):3282–3289, 1992.

[56] F Narberhaus, K Giebeler, and H Bahl. Molecular characterization of the *dnaK* gene region of *Clostridium acetobutylicum*, including *grpE*, *dnaJ*, and a new heat shock gene. *Journal of bacteriology*, 174(10):3290–3299, 1992.

[57] Carlos J Paredes, Keith V Alsaker, and Eleftherios T Papoutsakis. A comparative genomic view of *Clostridial* sporulation and physiology. *Nature Reviews Microbiology*, 3(12):969–978, 2005.

[58] 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.

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

[60] Wolfgang Schumann. The *Bacillus subtilis* heat shock stimulon. *Cell stress & chaperones*, 8(3):207, 2003.

[61] Xian-Zhi Li and Hiroshi Nikaido. Efflux-mediated drug resistance in bacteria. *Drugs*, 69(12):1555–1623, 2009.

[62] Bin-Bing S Zhou and Stephen J Elledge. The dna damage response: putting checkpoints in perspective. *Nature*, 408(6811):433–439, 2000.

[63] CF Beck, R Mutzel, J Barbe, and W Müller. A multifunctional gene (tetr) controls tn10-encoded tetracycline resistance. *Journal of bacteriology*, 150(2):633–642, 1982.

[64] Xian-Zhi Li, Li Zhang, and Keith Poole. Role of the multidrug efflux systems of *Pseudomonas aeruginosa* in organic solvent tolerance. *Journal of bacteriology*, 180(11):2987–2991, 1998.

[65] Michael M Cox and John R Battista. *Deinococcus radioduransâĂŤthe consummate survivor*. *Nature Reviews Microbiology*, 3(11):882–892, 2005.

[66] Xu Sun, Zileena Zahir, Karlene H Lynch, and Jonathan J Dennis. An antirepressor, srpr, is involved in transcriptional regulation of the srpabc solvent tolerance efflux pump of *Pseudomonas putida* s12. *Journal of bacteriology*, 193(11):2717–2725, 2011.

[67] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[68] 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.

[69] 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.

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

[71] 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.

[72] Andreas Pich, Franz Narberhaus, and Hubert Bahl. Induction of heat shock proteins during initiation of solvent formation in *Clostridium acetobutylicum*. *Applied Microbiology and Biotechnology*, 33(6):697–704, 1990.

[73] Joseph S Terracciano, Eliezer Rapaport, and Eva R Kashket. Stress-and growth phase-associated proteins of *Clostridium acetobutylicum*. *Applied and environmental microbiology*, 54(8):1989–1995, 1988.

[74] Davide Roncarati, Alberto Danielli, and Vincenzo Scarlato. The hrca repressor is the thermosensor of the heat-shock regulatory circuit in the human pathogen *Helicobacter pylori*. *Molecular microbiology*, 92(5):910–920, 2014.

[75] Elke Krüger, Tarek Msadek, and Michael Hecker. Alternate promoters direct stress-induced transcription of the *Bacillus subtilis clpC* operon. *Molecular microbiology*, 20(4):713–723, 1996.

[76] Haike Antelmann, Roland Schmid, and Michael Hecker. The nad synthetase nade (outb) of *Bacillus subtilis* is a σb-dependent general stress protein. *FEMS microbiology letters*, 153(2):405–409, 1997.

[77] Christian Scharf, Sabine Riethdorf, Henrik Ernst, Susanne Engelmann, Uwe Völker, and Michael Hecker. Thioredoxin is an essential protein induced by multiple stresses in *Bacillus subtilis*. *Journal of bacteriology*, 180(7):1869–1877, 1998.

[78] Ulf Gerth, Elke Krüger, Isabelle Derré, Tarek Msadek, and Michael Hecker. Stress induction of the *Bacillus subtilis clpP* gene encoding a homologue of the

proteolytic component of the clp protease and the involvement of clpp and clpx in stress tolerance. *Molecular microbiology*, 28(4):787–802, 1998.

[79] Soraya Chaturongakul, Sarita Raengpradub, M Elizabeth Palmer, Teresa M Bergholz, Renato H Orsi, Yuewei Hu, Juliane Ollinger, Martin Wiedmann, and Kathryn J Boor. Transcriptomic and phenotypic analyses identify coregulated, overlapping regulons among prfa, ctsr, hrca, and the alternative sigma factors σb, σc, σh, and σl in *Listeria monocytogenes*. *Applied and environmental microbiology*, 77(1):187–200, 2011.

[80] Elke Krüger and Michael Hecker. The first gene of the *Bacillus subtilis clpC* operon, *ctsR*, encodes a negative regulator of its own operon and other class iii heat shock genes. *Journal of bacteriology*, 180(24):6681–6688, 1998.

[81] Jakob Fuhrmann, Andreas Schmidt, Silvia Spiess, Anita Lehner, Kürşad Turgay, Karl Mechtler, Emmanuelle Charpentier, and Tim Clausen. Mcsb is a protein arginine kinase that phosphorylates and inhibits the heat-shock regulator ctsr. *Science*, 324(5932):1323–1327, 2009.

[82] Liang Tao, Partho Chattoraj, and Indranil Biswas. Ctsr regulation in *mcsAB*-deficient gram-positive bacteria. *Journal of bacteriology*, 194(6):1361–1368, 2012.

[83] AKW Elsholz, K Hempel, S Michalik, K Gronau, D Becher, M Hecker, and U Gerth. Activity control of the clpc adaptor mcsb in *Bacillus subtilis*. *Journal of bacteriology*, 193(15):3887–3893, 2011.

[84] S Leininger, T Urich, M Schloter, L Schwark, J Qi, GW Nicol, JI Prosser, SC Schuster, and C Schleper. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, 442(7104):806–809, 2006.

[85] Tse-Wen Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods*, 65(1):217–223, 1983.

[86] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, 2000.

[87] Vishwanath R Iyer, Christine E Horak, Charles S Scafe, David Botstein, Michael Snyder, and Patrick O Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409(6819):533–538, 2001.

[88] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[89] Alexandra Sittka, Sacha Lucchini, Kai Papenfort, Cynthia M Sharma, Katarzyna Rolle, Tim T Binnewies, Jay CD Hinton, and Jörg Vogel. Deep sequencing analysis of small noncoding rna and mrna targets of the global post-transcriptional regulator, hfq. *PLoS genetics*, 4(8):e1000163, 2008.

[90] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. Hits-clip yields genome-wide insights into brain alternative rna processing. *Nature*, 456(7221):464–469, 2008.

[91] Sabarinath Jayaseelan, Francis Doyle, and Scott A Tenenbaum. Profiling post-transcriptionally networked mrna subsets using rip-chip and rip-seq. *Methods*, 67(1):13–19, 2014.

[92] Stanimir S Ivanov, Alicia S Chung, Zheng-long Yuan, Ying-jie Guan, Katherine V Sachs, Jonathan S Reichner, and Y Eugene Chin. Antibodies immobilized as arrays to profile protein post-translational modifications in mammalian cells. *Molecular & Cellular Proteomics*, 3(8):788–795, 2004.

[93] David G Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998.

[94] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[95] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, et al. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, 2007.

[96] Stephan C Schuster. Next-generation sequencing transforms todayâĂŹs biology. *Nature*, 200(8), 2007.

[97] Karen M Wassarman, Francis Repoila, Carsten Rosenow, Gisela Storz, and Susan Gottesman. Identification of novel small rnas using comparative genomics and microarrays. *Genes & development*, 15(13):1637–1651, 2001.

[98] Alexei Aravin, Dimos Gaidatzis, Sébastien Pfeffer, Mariana Lagos-Quintana, Pablo Landgraf, Nicola Iovino, Patricia Morris, Michael J Brownstein, Satomi Kuramochi-Miyagawa, Toru Nakano, et al. A novel class of small rnas bind to mili protein in mouse testes. *Nature*, 442(7099):203–207, 2006.

[99] Cheng Lu, Blake C Meyers, and Pamela J Green. Construction of small rna cdna libraries for deep sequencing. *Methods*, 43(2):110–117, 2007.

[100] Ho-Ching Tiffany Tsui, Hon-Chiu Eastwood Leung, and Malcolm E Winkler. Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in escherichia coli k-12. *Molecular microbiology*, 13(1):35–49, 1994.

[101] JACLYN SHINGARA, KERRI KEIGER, JEFFREY SHELTON, WALAIRAT LAOSINCHAI-WOLF, PATRICIA POWERS, RICHARD CONRAD, DAVID BROWN, and EMMANUEL LABOURIER. An optimized isolation and labeling platform for accurate microrna expression profiling. *Rna*, 11(9):1461–1470, 2005.

[102] Henk PJ Buermans, Yavuz Ariyurek, Gertjan van Ommen, Johan T den Dunnen, and Peter AC't Hoen. New methods for next generation sequencing based microrna expression profiling. *BMC genomics*, 11(1):716, 2010.

[103] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.

[104] Arthur L Delcher, Douglas Harmon, Simon Kasif, Owen White, and Steven L Salzberg. Improved microbial gene identification with glimmer. *Nucleic acids research*, 27(23):4636–4641, 1999.

[105] Mario Stanke and Burkhard Morgenstern. Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, 33(suppl 2):W465–W467, 2005.

[106] Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, et al. The rast server: rapid annotations using subsystems technology. *BMC genomics*, 9(1):75, 2008.

[107] Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.

[108] Pierre Nicolas, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, Elodie Marchadier, Mark Hoebeke,

Stéphane Aymerich, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*, 335(6072):1103–1106, 2012.

[109] Kayoko Yamada, Jun Lim, Joseph M Dale, Huaming Chen, Paul Shinn, Curtis J Palm, Audrey M Southwick, Hank C Wu, Christopher Kim, Michelle Nguyen, et al. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, 302(5646):842–846, 2003.

[110] X Shirley Liu. Getting started in tiling microarray analysis. *PLoS computational biology*, 3(10):e183, 2007.

[111] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. *Nature methods*, 8(6):469–477, 2011.

[112] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[113] Julia Berretta and Antonin Morillon. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO reports*, 10(9):973–982, 2009.

[114] Jana Nejepinska, Radek Malik, Martin Moravec, and Petr Svoboda. Deep sequencing reveals complex spurious transcription from transiently transfected plasmids. *PloS one*, 7(8):e43283, 2012.

[115] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[116] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[117] Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, John A Stamatoyannopoulos, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[118] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

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

[120] Karla D Passalacqua, Anjana Varadarajan, Brian D Ondov, David T Okou, Michael E Zwick, and Nicholas H Bergman. Structure and complexity of a bacterial transcriptome. *Journal of bacteriology*, 191(10):3203–3211, 2009.

[121] Hedda Høvik, Wen-Han Yu, Ingar Olsen, and Tsute Chen. Comprehensive transcriptome analysis of the periodontopathogenic bacterium *Porphyromonas gingivalis* w83. *Journal of bacteriology*, 194(1):100–114, 2012.

[122] Timothy T Perkins, Robert A Kingsley, Maria C Fookes, Paul P Gardner, Keith D James, Lu Yu, Samuel A Assefa, Miao He, Nicholas J Croucher, Derek J Pickard, et al. A strand-specific rna-seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS genetics*, 5(7):e1000569, 2009.

[123] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131, 2010.

[124] Iwanka Kozarewa, Zemin Ning, Michael A Quail, Mandy J Sanders, Matthew Berriman, and Daniel J Turner. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+ c)-biased genomes. *Nature methods*, 6(4):291–295, 2009.

[125] Alec W. Picard, 2009.

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

[127] Brian J Haas, Melissa Chin, Chad Nusbaum, Bruce W Birren, and Jonathan Livny. How deep is deep enough for rna-seq profiling of bacterial transcriptomes? *BMC genomics*, 13(1):734, 2012.

[128] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[129] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment

(miame)âĂŤtoward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

[130] Jun Li, Hui Jiang, and W Wong. Method modeling non-uniformity in short-read rates in rna-seq data. *Genome Biol*, 11(5):R25, 2010.

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

[132] 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.

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

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

[135] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[136] W James Kent. BlatâĂŤthe blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[137] Carlos J Paredes, Isidore Rigoutsos, and E Terry Papoutsakis. Transcriptional organization of the clostridium acetobutylicum genome. *Nucleic acids research*, 32(6):1973–1981, 2004.

[138] Nicolas Sierro, Yuko Makita, Michiel de Hoon, and Kenta Nakai. Dbtbs: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic acids research*, 36(suppl 1):D93–D96, 2008.

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

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

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

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

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

[144] Jörg Sander, Xuejie Qin, Zhiyong Lu, Nan Niu, and Alex Kovarsky. Automatic extraction of clusters from hierarchical clustering representations. In *Advances in Knowledge Discovery and Data Mining*, pages 75–87. Springer, 2003.

[145] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[146] 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.

[147] Shawn W Jones, Carlos J Paredes, Bryan Tracy, Nathan Cheng, Ryan Sillers, Ryan S Senger, and Eleftherios T Papoutsakis. The transcriptional program underlying the physiology of *Clostridial* sporulation. *Genome Biol*, 9(7):R114, 2008.

[148] Ming Hu, Yu Zhu, Jeremy MG Taylor, Jun S Liu, and Zhaohui S Qin. Using poisson mixed-effects model to quantify transcript-level gene expression in rna-seq. *Bioinformatics*, 28(1):63–68, 2012.

[149] Timothy Daley and Andrew D Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–327, 2013.

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

[151] Thermo Scientific. *260/280 and 260/230 Ratios*, April 2008.

[152] Agilent Technologies. *Agilent 2100 bioanalyzer. Application Compendium. 2007. Agilent Technologies, 156pp*, June 2007. Printed in Germany.

[153] Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & development*, 23(12):1379–1386, 2009.

[154] Lichun Jiang, Felix Schlesinger, Carrie A Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for rna-seq experiments. *Genome research*, 21(9):1543–1551, 2011.

[155] Sándor J Piros and Ghaleb A Husseini. Preliminary modeling of transfer rna kinetics in the cytoplasm of *Escherichia coli* bacteria. *Advanced Science Letters*, 3(1):28–36, 2010.

[156] Jonathan A Bernstein, Arkady B Khodursky, Pei-Hsun Lin, Sue Lin-Chao, and Stanley N Cohen. Global analysis of mrna decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent dna microarrays. *Proceedings of the National Academy of Sciences*, 99(15):9697–9702, 2002.

[157] Ramesh V Nair, Edward M Green, David E Watson, George N Bennett, and Eleftherios T Papoutsakis. Regulation of the *sol* locus genes for butanol and acetone formation in *Clostridium acetobutylicum* atcc 824 by a putative transcriptional repressor. *Journal of bacteriology*, 181(1):319–330, 1999.

[158] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[159] Maureen J Donlin. Using the generic genome browser (gbrowse). *Current Protocols in Bioinformatics*, pages 9–9, 2009.

[160] Ulrike Gerischer and Peter Dürre. mrna analysis of the adc gene region of *Clostridium acetobutylicum* during the shift to solventogenesis. *Journal of bacteriology*, 174(2):426–433, 1992.

[161] Uwe Sauer and Peter Dürre. Differential induction of genes related to solvent formation during the shift from acidogenesis to solventogenesis in continuous culture of *Clostridium acetobutylicum*. *FEMS microbiology letters*, 125(1):115–120, 1995.

[162] Richard W Welch, Frederick B Rudolph, and E Terry Papoutsakis. Purification and characterization of the nadh-dependent butanol dehydrogenase from *Clostridium acetobutylicum* atcc 824. *Archives of biochemistry and biophysics*, 273(2):309–318, 1989.

[163] Peter Dürre, Anita Kuhn, Matthias Gottwald, and Gerhard Gottschalk. Enzymatic investigations on butanol dehydrogenase and butyraldehyde dehydrogenase in extracts of *Clostridium acetobutylicum*. *Applied microbiology and biotechnology*, 26(3):268–272, 1987.

[164] Karen Montzka Wassarman and Gisela Storz. 6s rna regulates *E. coli* rna polymerase activity. *Cell*, 101(6):613–623, 2000.

[165] Georg Homuth, Axel Mogk, and Wolfgang Schumann. Post-transcriptional regulation of the *Bacillus subtilis dnaK* operon. *Molecular microbiology*, 32(6):1183–1197, 1999.

[166] Hubert Bahl, Harald Müller, Susanne Behrens, Heinke Joseph, and Franz Narberhaus. Expression of heat shock genes in *Clostridium acetobutylicum*. *FEMS microbiology reviews*, 17(3):341–348, 1995.

[167] Latonia M Harris, Neil E Welker, and Eleftherios T Papoutsakis. Northern, morphological, and fermentation analysis of spo0a inactivation and overexpression in clostridium acetobutylicum atcc 824. *Journal of bacteriology*, 184(13):3586–3597, 2002.

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

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

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

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

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

[173] 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.

[174] Donghyuk Kim, Jay Sung-Joong Hong, Yu Qiu, Harish Nagarajan, Joo-Hyun Seo, Byung-Kwan Cho, Shih-Feng Tsai, and Bernhard ÃŸ 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.

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

[176] 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.

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

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

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

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

[181] 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.

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

[183] 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.

[184] 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.

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

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

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

[188] 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.

[189] 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.

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

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

[192] Yong Zhang, Tao Liu, Clifford A Meyer, JÃľrÃťme 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.

[193] 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.

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

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

[196] 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.

[197] Anja Schulz and Wolfgang Schumann. *hrcA*, the first gene of the *Bacillus subtilis dnaK* operon encodes a negative regulator of class i heat shock genes. *Journal of bacteriology*, 178(4):1088–1093, 1996.

[198] Simone Fleige and Michael W Pfaffl. Rna integrity and the effect on the real-time qrt-pcr performance. *Molecular aspects of medicine*, 27(2):126–139, 2006.

[199] Subramanian S Ajay, Stephen CJ Parker, Hatice Ozel Abaan, Karin V Fuentes Fajardo, and Elliott H Margulies. Accurate and comprehensive sequencing of personal genomes. *Genome research*, 21(9):1498–1505, 2011.

[200] Uwe Sauer, Anke Treuner, Malte Buchholz, Joseph D Santangelo, and P Dürre. Sporulation and primary sigma factor homologous genes in *Clostridium acetobutylicum*. *Journal of bacteriology*, 176(21):6572–6582, 1994.

[201] Tse-Wen Chang. Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods*, 65(1):217–223, 1983.

[202] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[203] Alexander KW Elsholz, Kristina Hempel, Dierk-Christoph Pöther, Dörte Becher, Michael Hecker, and Ulf Gerth. Ctsr inactivation during thiol-specific stress in low gc, gram+ bacteria. *Molecular microbiology*, 79(3):772–785, 2011.

[204] Shen Mynn Tan, Rory Kirchner, Jingmin Jin, Oliver Hofmann, Larry McReynolds, Winston Hide, and Judy Lieberman. Sequencing of captive target transcripts identifies the network of regulated genes and functions of primate-specific mir-522. *Cell reports*, 8(4):1225–1239, 2014.

[205] Rikizo Aono, Norihiko Tsukagoshi, and Mami Yamamoto. Involvement of outer membrane protein tolc, a possible member of the mar-sox regulon, in maintenance and improvement of organic solvent tolerance of *Escherichia coli K-12*. *Journal of bacteriology*, 180(4):938–944, 1998.

Appendix A

SUPPLEMENTARY FIGURES AND TABLES

Table A.1: Read Summary

ID	Stress	Time	Rep.	Barcode	Lane	Total	rRNA	non-rRNA	Mapped	Chrom.	pSol1	Proper-pairs
1	NS	15	A	5'-ATCACG-3'	1-3	48859456	30834543	18024913	17670942	16787882	883060	15125964
2	NS	150	A	5'-CGATGT-3'	1-3	45478896	29206366	16272530	15949307	15544957	404350	13541382
3	BuOH	15	A	5'-TTAGGC-3'	1-3	54023718	26877428	27146290	26274654	24625624	1649030	32258512
4	BuOH	150	A	5'-TGACCA-3'	1-3	57478096	20974221	36503875	35309409	34417748	891661	20252538
5	BA	15	A	5'-ACAGTG-3'	1-3	44806788	27899812	16906976	16122130	15438881	683249	15684020
6	BA	150	A	5'-GCCAAT-3'	1-3	50528992	31702977	18826015	18262777	17793342	469435	12777076
7	NS	15	A	5'-CAGATC-3'	1-3	48703316	28283201	20420115	19990266	19591705	398561	17710744
8	NS	150	A	5'-ACTTGA-3'	1-3	42733074	25975735	16757339	16286286	15309938	976348	14097826
9	BuOH	15	B	5'-GATCAG-3'	1-3	49506120	38595998	10910122	10630559	10040255	590304	8852576
10	BuOH	150	B	5'-TAGCTT-3'	1-3	43545678	27183105	16362573	16061439	15471863	589576	14039214
11	BA	15	B	5'-GGCTAC-3'	1-3	45916674	23845246	22071428	21676193	20634570	1041623	19140554
12	BA	150	B	5'-CTTGTA-3'	1-3	44134074	27588144	16545930	16249079	15922530	326549	14253616
13	NS	75	A.1	5'-AGTCAA-3'	1-3	54189730	33065951	21123779	20565436	19480660	1084776	19434268
14	NS	270	A.1	5'-AGTTCC-3'	1-3	52034630	28645744	23388886	22789276	22389726	399550	17582076
15	BuOH	75	A.1	5'-ATGTCA-3'	1-3	47572376	32667456	14904920	14470393	14090265	380128	17543516
16	BuOH	270	A.1	5'-CCGTCC-3'	1-3	55768414	32755408	23013006	22234995	22101359	133636	11907496
17	BA	75	A.1	5'-GTAGAG-3'	1-3	56830036	33780862	23049174	22569837	21724902	844935	19508768
18	BA	270	A.1	5'-GTGAAA-3'	1-3	46916380	28520421	18395959	18033076	17771930	261146	15819006
19	NS	75	A.2	5'-ATCACG-3'	4-5	55562942	34712882	20850060	20379057	19412583	966474	17470744
20	NS	270	A.2	5'-CGATGT-3'	4-5	58493836	36066923	22426913	22035846	21362793	673053	18936254
21	NS	75	B	5'-TTAGGC-3'	4-5	52771634	34117151	18654483	18246977	17845109	401868	15641712
22	NS	270	B	5'-TGACCA-3'	4-5	36899276	16765267	20134009	19170869	18885771	285098	13998218
23	BuOH	75	A.2	5'-ACAGTG-3'	4-5	51283910	25058052	26225858	25321716	24678137	643579	6212554
24	BuOH	270	A.2	5'-GCCAAT-3'	4-5	55827282	40650151	15177131	13930352	13896505	33847	21755582
25	BuOH	75	B	5'-CAGATC-3'	4-5	44530728	30358824	14171904	13802738	13587661	215077	17404278
26	BuOH	270	B	5'-ACTTGA-3'	4-5	44125876	22665534	21460342	20645400	20505901	139499	11705956
27	BA	75	A.2	5'-GATCAG-3'	4-5	53937782	41434722	12503060	12028549	11739925	288624	9387978
28	BA	270	A.2	5'-TAGCTT-3'	4-5	51836778	38673503	13163275	12701884	12581754	120130	9735850
29	BA	75	B	5'-GGCTAC-3'	4-5	57202968	42763499	14439469	13790811	13332328	458483	9259022
30	BA	270	B	5'-CTTGTA-3'	4-5	47920082	37290205	10629877	10221661	10020490	201171	7777560
Sum	N/A	N/A	N/A	N/A	N/A	1499419542	928959331	570460211	553421914	536987094	16434820	458814860
Avg	N/A	N/A	N/A	N/A	N/A	49980651	30965311	19015340	18447397	17899570	547827	15293829

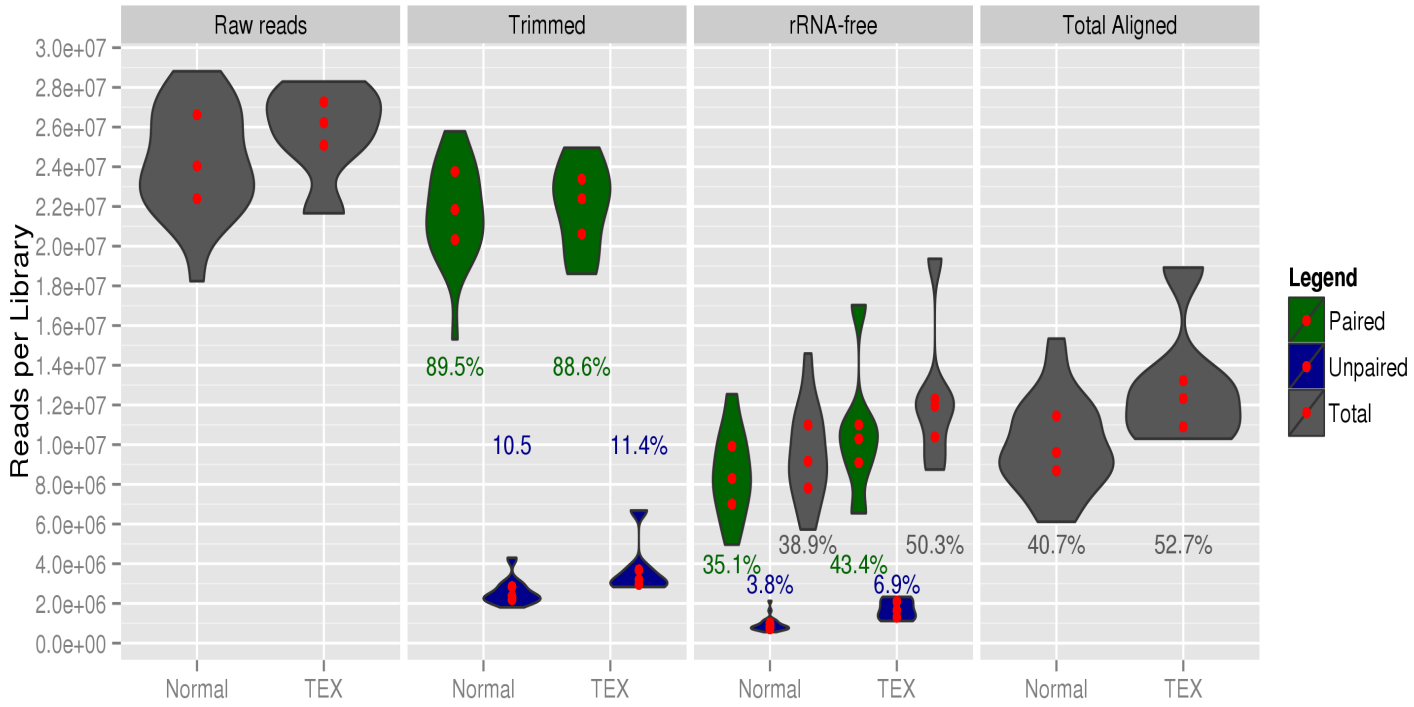


Figure A.1: Read Trimming, Filtering, and Alignment

Raw reads from each library were processed, resulting in both paired and unpaired reads. These reads were then aligned to rRNA sequences, with 97% of the remaining mRNA reads aligning to the genome.

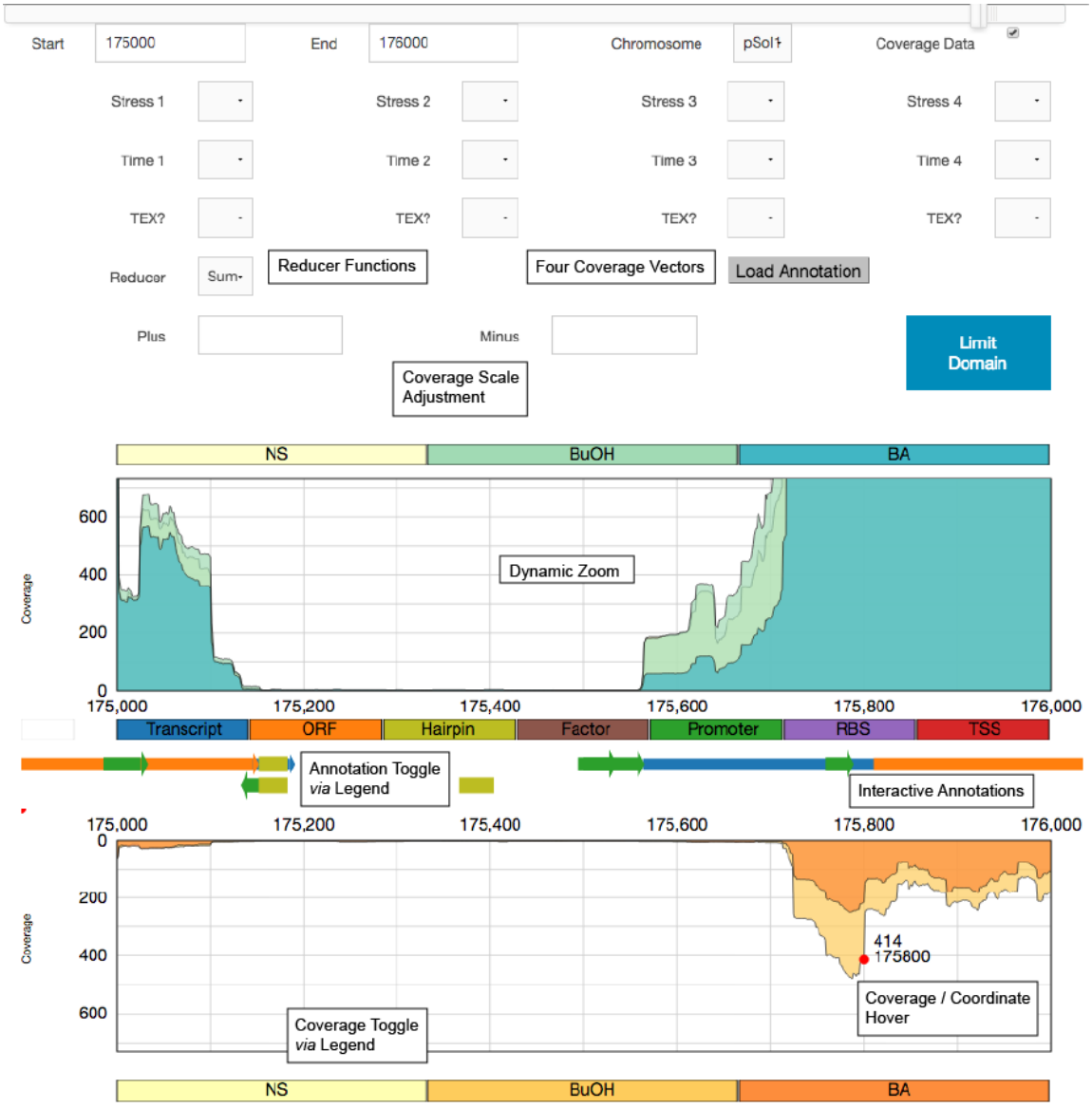


Figure A.2: Genome Browser
The customized genome browser has unique features that facilitate the exploration of the sequencing dataset at high resolution.

Appendix B

PERMISSION LETTERS

From: "Shawn Jones"

Subject: RE: Permission letter

To: "'Matthew Ralston'" <mrals@udel.edu>

Hi Matt,

You have my permission to use the figure. Congratulations on the thesis!

Shawn

From: Matthew Ralston [mailto:mrals@udel.edu]

To: Shawn Jones

Subject: Permission letter

Hi Shawn,

On page 31 in my thesis, I adapted a figure from your Genome Biology paper "The transcriptional program underlying the physiology of *Clostridial* sporulation." It is a beautiful growth curve and metabolite profile that was easy to adapt to illustrate the time points and metabolites that were part of this experimental design. I have attributed the figure to your paper (pg 31) but require a permission letter to formally associate that the figure is yours. Would you mind if I use the adapted growth curve as a figure in my thesis? The figure is included in a pdf accessible below.

Thank you,

Matt Ralston