

Papoutsakis Eleftherios (Orcid ID: 0000-0002-1077-1277)

## **RNAseq-based Transcriptome Assembly of *Clostridium acetobutylicum* for Functional Genome Annotation and Discovery**

Matthew T. Ralston<sup>2,3</sup> and Eleftherios T. Papoutsakis<sup>1,2\*</sup>

<sup>1</sup> Dept. of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE 19711

<sup>2</sup> Molecular Biotechnology Laboratory, Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711

<sup>3</sup> Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711

Running Head: Assembly of *C. acetobutylicum* RNAseq transcriptome

\* To whom correspondence should be addressed. Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711;  
Email: epaps@udel.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as doi: [10.1002/aic.16396](https://doi.org/10.1002/aic.16396)

**(Abstract)**

Accurate genome annotations are essential in modern biology and biotechnology, yet they are still largely based on genome sequencing and comparative analyses. We show that the *Clostridium acetobutylicum* genome annotation can be markedly improved by integrating bioinformatic predictions with RNA sequencing (RNAseq) data. Samples were acquired under butanol, butyrate, and unstressed treatments across various growth. Analysis of an initial assembly revealed errors due to background signals and limitations of assembly algorithms. Hurdles for RNAseq transcriptome mapping include optimizing library complexity and sequencing depth, yet most studies report low sequencing depth and ignore the effect of ribosomal RNA abundance. An integrative analysis was developed to combine motif predictions, single-nucleotide resolution sequencing depth, and library complexity to resolve difficulties in assembly curation. This minimized false positive error and determined gene boundaries, in some cases, to the exact basepair of prior studies. This is the first strand-specific transcriptome assembly in a *Clostridium* organism.

**Key Words:** RNAseq, novel transcripts, acetone-butanol, small RNAs, antisense RNAs

## Introduction

RNA sequencing (RNAseq) is a powerful method for microbial genomics, producing hundreds of millions of sequenced reads from RNA populations<sup>1,2</sup> roughly proportional to expression levels<sup>3</sup>. Basepair level resolution of an entire transcriptome allows researchers to map transcript boundaries<sup>4</sup> assemble full-length transcripts without reference genomes<sup>5</sup>, detect novel transcripts<sup>6</sup>, and abundant antisense regulation<sup>7</sup>. Use of this technique can improve genome annotations by organizing predicted open reading frames into experimentally verified transcriptional units (TUs)<sup>8</sup>. However, there is considerable heterogeneity in the methodology and results from early transcriptome mapping studies in bacteria<sup>4,9-12</sup>. While guidelines for many second-generation sequencing applications in mice and human have been established<sup>13-15</sup>, no such recommendations exist for prokaryotic projects. Additionally, differences in approaches employed in bacterial studies raise questions about experimental variables and analytical techniques, such as the appropriate sequencing depth for microbial transcriptomes<sup>11</sup>.

Several experimental factors affect the useful sequencing depth and limit the sensitivity of RNAseq to detect rare cDNAs from low abundance transcripts and transcript boundaries (e.g. transcription start sites (TSS)). First, most early transcriptome mapping studies in bacteria did not utilize strand-specific RNA sequencing techniques<sup>9,10,16-19</sup>. Strand-specific RNAseq data are useful if not essential for mapping dense prokaryotic genomes<sup>8</sup>. Confirmation of open reading frame (ORF) predictions and transcriptional unit organization is difficult to determine without knowing the strand and orientation of an active genomic region. Additionally, antisense transcriptional regulators (i.e. *cis*-antisense small RNAs) are impossible to identify without strand specific information. Next, ribosomal RNA removal rates are highly variable in the literature and depend on genomic GC content<sup>8,11,20,21</sup>. Many early studies have not reported the efficiency of rRNA removal<sup>9,10,16,18</sup>; when these statistics have been described, as much as 92% of sequenced reads were ribosomal after a hybridization-based rRNA removal

strategy<sup>19</sup>. Moreover, many studies have unique alignment rates below 50% as a result of sequencing errors and low library complexity.

Sequencing depth is a critical issue<sup>8,11,22-24</sup> affecting the false positive (type I) and false negative (type II) error rates in the data. For example, one study reported that 40% of the ORFs were not completely covered by reads<sup>12</sup>, suggesting that additional sequencing/sensitivity was required to reduce the false negative error rate for transcript and TSS discovery<sup>24</sup>. On the other hand, while it has been shown that deep sequencing (> 100M reads) is sufficiently sensitive to detect rare transcriptomic features<sup>14</sup>, sequencing to saturation may also detect spurious transcripts and low-level contaminants, increasing the false positive error rate. The appropriate sequencing depth depends on the level of sensitivity desired and the tolerance for false positive signals. Un-normalized depth alone is often used by researchers to estimate transcript boundaries<sup>9,10,16-19</sup>, with no discussion of local biases in the depth signal<sup>25,26</sup>, limit of detection<sup>12</sup>, or background signals<sup>11,27,28</sup>. False positive depth signal can occur from DNA contamination<sup>28</sup>, spurious transcription<sup>11,27</sup>, PCR over-amplification, and platform specific bias<sup>25,28</sup>. However, sequencing complexity is less prone to such artifacts at the depths used by most studies. The limit of detection for interesting transcripts also depends on library complexity and is more useful than sequencing depth for some objectives<sup>24</sup>.

Transcriptome assembly algorithms can reconstruct full-length transcripts, even without a reference genome (*de novo* assembly). *Ab initio* assembly methods utilize a reference sequence to reproducibly generate transcript boundaries. A paired-end, strand-specific library preparation produces the highest quality transcriptomic datasets<sup>8</sup>, although few assembly algorithms can leverage both paired-end and strand-specific data for *ab initio* assembly<sup>5</sup>.

A final challenge to transcriptome mapping is the evaluation and optimization of assembly results. Assembly quality is affected by sensitivity of the dataset and the presence of undesirable reads from spurious transcription or DNA contamination<sup>11,27,28</sup>.

Genome assembly quality is usually assessed by counts of contigs, singletons, and the distributions of their lengths, although these classical genomic metrics do not necessarily reflect optimal transcriptomic outcomes<sup>8,29</sup>. Instead, functional metrics such as the number of assembled reference ORFs<sup>30</sup> or similarity to a related organism's ORFs<sup>29</sup> are more relevant transcriptomic metrics to optimize.

Here we address these issues in a transcriptome assembly of *Clostridium acetobutylicum*, the model organism for the Acetone Butanol Ethanol (ABE) fermentation and the *Clostridium* sporulation program<sup>31-33</sup>. Genomic knowledge of this organism is largely the result of the initial sequencing of its genome in 2001<sup>34</sup>. The original annotation is based on ORF predictions and BLAST hits. It is clear now that this annotation requires improvements and validation through experimental data and methods, such as RNA-seq transcriptome mapping. While a few transcripts in *C. acetobutylicum* have been characterized<sup>34-39</sup>, for most of the ORFs, precise transcripts boundaries remain unknown.

## Materials and Methods

### *Microbial cultures with metabolite stress*

*C. acetobutylicum* ATCC 824 was cultured anaerobically in 4-L New Brunswick Scientific BioFlo 310 bioreactors at 37°C, pH  $\geq$  5.0, as described<sup>6,40</sup>. When the cultures were grown to  $A_{600}=1$ , the N<sub>2</sub> flow rate was decreased to 50 mL/min and cultures were either stressed to a final concentration of 60 mM *n*-butanol, 40 mM potassium butyrate, or left unstressed, as described. 15 mL samples were collected at 15, 75, 150, and 270 min after treatment. Samples were collected for RNAseq analysis<sup>6,40</sup>.

### *RNA preparation, quality assessment, & library preparation for RNAseq analysis.*

These were carried out as described<sup>6,40</sup>. Briefly, samples for RNA isolation were collected by centrifugation at 5000 rpm at 4°C for 10 min and the pellets stored at -80°C. RNA was isolated using the Qiagen's miRNeasy Mini kit. After RNA extraction, mRNA and sRNA were enriched by using Microbe Express kit from Ambion® kit. The Ovation Prokaryotic RNA-Seq System (NuGEN® Technologies, San Carlos, CA) was used to synthesize cDNA from 500 ng of enriched RNA as follows. 2  $\mu$  L of first primer mix was added to 500 ng of RNA and incubated at 65°C for 5 min. Then, 10  $\mu$ L of the master mix (first strand buffer and enzyme) were added for first strand synthesis followed by the purification of the first strand cDNA using the Qiagen QiaQuick PCR purification kit. The last step of cDNA synthesis is synthesis of the 2<sup>nd</sup> strand, which was then purified using the Minelute Reaction clean up kit (Qiagen) and eluted in 10  $\mu$ L of elution buffer. The elution buffer was used to make up the volume of the cDNA to 50  $\mu$ L. The resulting 50  $\mu$ L of cDNA was used to construct libraries using the TruSeq DNA Sample preparation kit (Illumina): cDNA underwent end repair, 3' end adenylation, adapter ligation and enrichment. AMPure XP beads were used to clean up the DNA fragments after each process. Library quality was checked using a Bioanalyzer before loading onto HiSeq 2000.

### ***Data processing***

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](https://github.com/MatthewRalston/NGS_scripts)). In brief, the fastq headers were briefly pre-processed for downstream applications by concatenating the two columns of the Casava 1.8+ header with an underscore. Then, the remaining sequencing adaptors were removed from the reads with Trimmomatic<sup>41</sup>, an algorithm that recognizes

and removes user-supplied adapter sequences. Base quality was 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 published *C. acetobutylicum* ATCC 824 genome<sup>34</sup> the data were subjected to *in silico* ribosomal RNA removal by aligning the reads to the rRNA sequences with Bowtie 2.1.0<sup>41</sup>. 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<sup>42</sup> and Picard (<http://broadinstitute.github.io/picard/>). 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.

### ***Transcriptome assembly and quality assessment***

Reference assembly was done with Trinity<sup>5</sup>. 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. Options for the algorithm can be found in the Github repository script trinity.sh.

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 and BLAT. It was determined that BLAST produced comparable and superior alignments in most cases; the BLAST alignment of assembled transcripts was used for further analysis. The alignment was converted 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 intergenic regions (IGRs) were identified and compared with both the reference annotation and according to the operon organization by Paredes *et al.*<sup>43</sup>. Custom Julia, Ruby, and R scripts were used to produce and compare assembly statistics. 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.

### ***Genome browser***

A genome browser was required for assembly curation and 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. The schema consisted of two simple tables, as no table joining was necessary for data retrieval. 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 customized query system was developed for increased speed, bypassing the ORM. 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<sup>44</sup>. 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 is described in<sup>45</sup>.

## Results

### *Ultra-deep sequencing of the *C. acetobutylicum* transcriptome as a basis for high-quality transcriptome assembly*

A fractional factorial experimental design was chosen to assess the expression profile of physiologically meaningful transcripts. The factors investigated were metabolite stress (60 mM butanol and 40 mM butyrate stress and unstressed control), time (15, 75, 150, 270 min post stress), and Terminator<sup>TM</sup>5' -phosphate dependent exonuclease (TEX) treatment<sup>40</sup> (for the enrichment of 5' end of RNAs containing Transcriptional Start Sites (TSS)), all assessed with duplicate biological replicates. The selected concentrations of metabolites induce a stress response that includes a variety of genes and programs<sup>6,40</sup>, yet allow for sufficient biomass production for generating libraries of sufficient complexity. The selected time points span a meaningful range of growth stages<sup>6</sup>, of relevance to the natural stress response and the transition to the stationary phase of culture<sup>46-50</sup>. In total, 30 RNA samples were enriched for primary transcripts and multiplexed across 5 lanes of the Illumina HiSeq 2500 platform.

High sequencing depth and library complexity are required for successful assembly<sup>24,29</sup>. Here, 1.5 billion reads were produced in a paired-end, 76-basepair configuration (750 million pairs). Comparison with related studies revealed that the unprecedented quantity of data produced was sufficient for assembly<sup>45</sup>. The 92% rRNA removal rate and 97% alignment rate indicated a strong overall data utilization rate. Over 30% of sequenced clusters (cumulatively 459 million reads) were properly paired (correct alignment orientation), thus generating enough information to cover the *C. acetobutylicum* transcriptome over 11,000-fold. These numbers exceed recommendations for transcript discovery projects in humans<sup>14</sup> and were sufficient for the objectives of this study. Additional figures and statistics regarding the sensitivity of the dataset can be found in<sup>45</sup>. This sensitive sequencing dataset was then used for transcriptome assembly and transcript boundary detection.

***Transcriptome assembly produces expected (canonical), but curation of the transcriptome assembly improves precision of genome assembly***

The sensitivity of the dataset suggested that precise estimates of transcript boundaries could be acquired, even for transcripts that might have low abundance. These boundaries can be produced by transcriptome assembly algorithms, which rely on the complexity of a sequencing dataset to generate overlap layout consensus (OLC) or de Bruijn graphs for traversal and transcript reconstruction<sup>8</sup>. These algorithms find the “least common denominator” transcript sequence among reads and have been extensively reviewed<sup>8</sup>. The Trinity assembly algorithm<sup>5</sup> was selected for its strand-specific, paired-end, and jaccard clipping features.

Two subsets of the data were first chosen for assembly and comparison. Most of the aligned reads (459M, 83%) were uniquely aligned and properly-paired; the remaining reads were single reads, duplicates, discordant pairs, or split pairs. It was hypothesized that these extra reads would produce a larger and more inclusive assembly than the smaller, high-quality subset of the proper pairs alone. Instead, the extra reads confounded the assembly graphs and resulted in a higher degree of misassembly. Thus, the properly paired subset produced an assembly that was selected as the optimal initial approximation of transcript boundaries.

In many cases, the initial assembly produced estimates of transcript boundaries that are consistent with previous literature values. For example, *groES* and *groEL* are solvent responsive class I heat shock genes<sup>49,51-54</sup> transcribed in a bicistronic operon on the chromosome of *C. acetobutylicum* (Figure 1). The assembled transcription start site closely matched the coordinate from a primer extension study<sup>39</sup>. The assembled transcription stop site was also consistent with a Rho-independent terminator, producing a transcript size of 2,131bp in agreement with an observed 2.2 kb band<sup>39</sup>. In this example, the assembled transcript did not require curation to improve the precision of boundary identification.

On the other hand, many assembled transcripts required some curation to better reflect genomic motifs and other signals from the RNAseq data. A manual curation method was used to improve the precision of the results and correct misassemblies. Curation was facilitated by a custom genome browser with optimized data density <sup>45</sup> (URL: <https://clostridia.dbi.udel.edu/CBrowser>, for details see Supporting Document SD1: Details for accessing the genome browser). In brief, assembled transcripts were examined for consistent patterns of depth with respect to reference CDSes (term used as a plural to CDS, **CoDing Sequence**), predicted Rho-independent terminators, and promoter motifs. This browser is largely useful for inspecting the stress and non-stress data used for the assembly and for inspecting the molecular genome elements, including TSSs, promoters, terminators, novel transcripts, sRNAs, and asRNAs. Some of these elements are discussed below. Several examples of the curation process can be found in <sup>45</sup>.

As a first example of the manual curation process, we detail the assembly of transcripts in the *sol* locus, which initially produced misassemblies, thus illustrating the challenges associated with deep sequencing background signals. The *sol* locus is coded on the pSOL1 megaplasmid of *C. acetobutylicum* <sup>55</sup> and consists of the tricistronic *sol* operon (*adhE1* (or *aad*)-*ctfA*-*ctfB*) <sup>56,57</sup> coded on the positive strand, and the convergent *adc* gene coded on the opposite strand. The bifunctional aldehyde-alcohol dehydrogenase (AdhE1, also known as AAD), coded by the *adhE1* (or *aad*) gene (CA\_P0162), is responsible for butanol production <sup>58</sup>, while the *ctfA* (CA\_P0163) and *ctfB* (CA\_P0164) genes code for the two subunits of the CoA-transferase (CoAT) that catalyzes the conversion of acetoacetyl-CoA to acetoacetate while transferring the CoA moiety to either acetate or butyrate <sup>59</sup>. Acetoacetate is converted to acetone upon the action of the *adc* (CA\_P0165)-coded acetoacetate decarboxylase (AADC) <sup>60,61</sup>. This locus also contains a small regulatory RNA(SolB)<sup>62</sup> upstream of (CA\_P0162), and a likely relevant or regulatory protein (CA\_P0161) <sup>63,64</sup>. The assembled *sol* operon transcript has a distal TSS identical to the location reported previously corresponding to a weak distal promoter

<sup>35,38</sup> (Figure 2a). Also observed was an increase in coverage consistent with the previously described proximal TSS corresponding to a strong proximal promoter <sup>35,38</sup>. Northern blots probing *ctfB* <sup>37</sup> and *adhE1* <sup>37,65</sup> indicate that the *sol* transcript is ca. 4 kb in length, terminating near a bifunctional Rho-independent terminator. Depth of coverage decreases after the end of the *sol* operon near this terminator, but background antisense signal [typically 1-5% <sup>66</sup>] from the *adc* gene was sufficiently complex for the assembly algorithm to misinterpret as true signal (Figure 2b). Correction of these artifacts through a curation process made the results consistent with previous results from a number of loci <sup>45</sup>.

As a result of detailed curation, the length of standard and novel transcripts displayed improved characteristics compared to their uncured counterparts (Figure 3). The distribution of transcript lengths agrees with the average transcript length of 1.1kb in *E. coli* (Figure 3)<sup>67</sup>. Untranslated region (UTR) lengths IGR lengths were similarly improved (Supporting Information Figures SF1 & SF2)<sup>67</sup>. In summary, 2,258 transcripts were detected. Of these, 1,635 canonical transcripts included 3,152 reference CDSes (83%) within their boundaries, accounting for much of the existing knowledge on this organism. Of the 627 remaining transcripts, 26 overlapped with previously described small RNAs <sup>6,68</sup>, leaving 601 as novel transcripts. The novel transcripts were smaller on average (Supporting Information Figure SF3), but with comparable coverage profiles. These novel transcripts thus represent additional non-coding RNAs or small protein coding transcripts. Of these novel transcripts, 452 were found to be at least partially antisense to another transcript and some of these are likely *cis* small RNAs (Supporting Information Excel File SX1). Many of these antisense transcripts were observed in physiologically or metabolically meaningful loci and a few select ones are discussed below.

### ***Novel transcriptional start sites, operon organization & transcripts***

Transcriptional start sites facilitate the investigation of the molecular basis of large physiological programs, and are thus important features of interest for transcriptome-mapping studies. These features facilitate future investigations of promoters and regulatory regions upstream of these sites. The regulatory complexity of the physiological program of this organism is very high<sup>31,40,54</sup>, probably due to the fact that soil organisms have been exposed to and respond to a broad range of environmental signals. Many of these complex interactions and the associated programs are orchestrated by the large number of sigma and transcription factors of this organism<sup>31,43</sup>. Knowledge of the coordinates of TSSs is the first step towards unraveling the complex cascading<sup>31,54</sup> and dense overlapping regulons<sup>43</sup> of the *C. acetobutylicum* sporulation, stress response and other environmentally responsive programs. Over 2,000 transcription start sites were discovered from this assembly, several of them novel (Supporting Information Excel File SX1).

In a first example, the pre-spore specific sigma factor F (SigF) is an alternative sigma factor that is responsible for initiating spore development. It is governed by a complex regulatory system involving at least one anti-sigma factor and one anti-anti sigma factor<sup>31,32</sup>. The *sigF* gene (CA\_C2306) is co-transcribed with an anti-sigma factor (CA\_C2307) and an anti-anti sigma factor (CA\_C2308) in an operon (Figure 4). The transcription start site reported here at base 2,411,363 on the chromosome is useful for understanding the complex regulation of the *Clostridia* sporulation pattern. Interestingly, our analysis has also revealed a non-trivial signal in an antisense orientation to the *sigF* operon that will be discussed in the next section.

A second example is the discovery of a new TSS, promoter and transcript from the locus of the *sol* operon discussed above. As discussed above, the *sol* operon has been assumed heretofore to generate one long 4-kb transcript (3973-bp precisely, based on the genome sequence and operon organization). However, our RNAseq data revealed a

sustained low coverage of read counts at the junction between the *adhE1* and *ctfA* transcripts (Figure 5). Based on this finding, we identified a previously unknown, putative terminator sequence within the first 60 bps of the *ctfA* coding sequence (coordinates 178488 to 178525 on the pSOL1 megaplasmid DNA). This Rho-independent terminator, which has a  $\Delta G$  of -9.6 kcal/mol, had not been identified previously due to its location within a coding region, but also for lack of *ctfA*-probing Northern blots until recently<sup>62</sup>. These recent blots identified extra bands of mRNAs deriving from the *sol* operon, two of them containing the *ctfA* transcript<sup>62</sup>. This prompted a search that led to the identification of a strong clostridial promoter, TTCATA(13)TATAAT, just upstream of the RBS of the *ctfA* gene, thus suggesting that the *ctfA-ctfB* genes can be independently expressed from such a promoter. As a result of these findings, there are a total of three potential transcripts from the *sol* operon: the long *adhE1(aad)-ctfA-ctfB* transcript, the *adhE1(aad)* transcript and the *ctfA-ctfB* transcript.

The RNAseq data also discovered novel genes. A notable one is the the CA\_C2079 gene. Slightly downstream of the *spo0A* gene (CA\_C2071) on the negative DNA strand, modest RNA expression from a long operon (CA\_C2078 to 2073) is followed by a “gap” DNA region of high expression. The coverage pattern of this region<sup>45</sup> is distinct from neighboring regions and corresponds to a putative protein (CA\_C2079) that is missing from the databases. This gene, which shares homology to efflux transporters, was originally included in the genome annotation, but has been removed from the data bases (e.g., KEGG) over the last few years. This is the first experimental evidence for its expression.

### ***Small antisense RNAs***

Natural antisense RNAs have been described for a number of bacteria<sup>7,28</sup> with some studies reporting thousands of antisense candidates in *E. coli*<sup>7</sup>. An antisense RNA regulating the glutamine synthetase expression has been reported in *C. acetobutylicum*, the first of its kind in this organism<sup>69</sup>. The large number of antisense transcripts observed

in other organisms were also observed in the stress responsive transcriptome of *C. acetobutylicum*. Over 400 possible antisense interactions were revealed from the integrative assembly (Supporting Information Excel File SX2). These antisense transcripts may regulate a large number of physiologically important systems in *C. acetobutylicum*, providing insight into the complex regulation of programs such as the *sigF* locus. These antisense small RNAs could be useful tools for metabolic engineering and biofuels research in this genus. We discuss two interesting examples of such asRNAs below.

To add to the complexity of the previously described *sigF* locus, opposite of this operon is an antisense RNA (Figure 4) that displays an alternate stress response profile. This difference between the stress response of the *sigF* operon and the antisense signal suggests separate regulatory mechanisms and functional roles. The lack of correlation between the per-base sequencing depth profile across this region combined with the level of expression (> 1-3% of sense expression) suggests that the observed antisense transcript is not an artifact of background antisense signal from library preparation. This novel antisense RNA transcript, is the first reported antisense regulator of the *sigF* locus. The sporulation program is regulated at multiple molecular levels and the presence of antisense transcription in the regulation of an important *Clostridium* program demonstrates the pervasive nature of antisense regulation.

The genus *Clostridium* possesses a number of hydrogen producing bacteria, which utilize available high-energy electrons (typically from reduced ferredoxin) to generate H<sub>2</sub> gas using hydrogenase enzymes. One of these hydrogenases, *hydA*, displayed antisense transcriptional activity under the conditions examined (Figure 6). The antisense patterns have strong expression levels over the background signal, are not strongly correlated with per-base sequencing depth levels on the sense strand, and display different stress response profiles than the *hydA* gene. Natural *hydA* antisense genes have not been

described in *C. acetobutylicum*, though artificial *hydA* antisense has been useful for understanding TNT bioremediation in this strain <sup>70</sup>.

Another example of antisense transcription concerns antisense signals in the locus coding for one of the most important operons, the BCS operon (CA\_C2712, 2711, 2710, 2709 and 2708), of this organism coding for the enzymes for the biosynthesis of butyryl-CoA from acetyl-CoA <sup>71,72</sup>. The operon codes for a 3-hydroxybutyryl-CoA dehydratase, butyryl-CoA dehydrogenase, electron transfer flavoproteins *eftA* and *eftB*, and a 3-hydroxybutyryl-CoA dehydrogenase. The locus of this operon also codes for two antisense RNAs (Figure 7) where the expression pattern does not reflect the transcriptional profiles for the BCS operon. This would seem to suggest that these putative antisense transcripts are also regulated differently from the BCS operon.

## Discussion

Previous sequencing studies focused on small portions of the transcriptome <sup>6,19</sup> and/or were limited by low sequencing depths <sup>6,10,17-19</sup> and non-directional approaches <sup>6,10,17-19</sup>. The mapping of this transcriptome represents the first strand specific mapping effort in a *Clostridium* organism, producing results that were consistent with previously described motifs and transcript boundaries. Additionally, functional and global metrics were consistent with findings in the well-characterized gram-negative *E. coli* <sup>67</sup>.

Genomic knowledge for any organism is a moving target, with far more conditions and genes to assess than resources to detect them. With firm evidence for 83% of reference ORFs under these conditions in addition to 627 novel transcripts, these findings will be useful for characterization and engineering research in this genus. Future work could include clustering of genes by expression profiles and determination of motifs associated with each cluster. Additionally, re-annotation of the genome or annotation of the novel transcripts could reveal novel ORFs, metabolic functions, signaling pathways and more. This work demonstrates the role of ‘omics measurements in improving

knowledge of microbial genomes and the opportunity to discover novel transcripts, proteins, and genes. At the same time, the challenges encountered here suggest the need for a discussion of depth requirements and improved assembly assessment methods.

Currently, assembly algorithms rely on sequencing complexity alone to determine if a transcribed region is significantly active. The non-uniform distribution of sequencing depth throughout the genome makes inferences based on thresholds unreliable (i.e. setting a minimum read threshold to determine active regions)<sup>6</sup>. To correct false positive errors, an integrative curation method was used to combine multiple genomic and transcriptomic signals to improve the precision of boundary estimates. In the future, improved assembly algorithms (perhaps based on unsupervised machine learning) incorporating these various signals may offer improved precision for deep transcriptome mapping.

**Acknowledgment**

This work was supported in part by a National Science Foundation (USA) grant (Award No. CBET-1511660) and a grant from the Department of Energy (USA, grant # DE-SC0007092). The funding agencies had no role in the experimental design, data collection and interpretation, or the decision to publish this research.

## Literature Cited

1. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456:53-59.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009;10:57-63.
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*. 2008;18:1509-1517.
4. Nicolas P, Mäder U, Dervyn E, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science*. 2012;335:1103-1106.
5. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*. 2011;29:644-652.
6. Venkataramanan KP, Jones SW, McCormick KP, et al. The *Clostridium* small RNome that responds to stress: the paradigm and importance of toxic metabolite stress in *C. acetobutylicum*. *BMC genomics*. 2013;14:849.
7. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT. Widespread Antisense Transcription in *Escherichia coli*. *Mbio*. 2010;1:1-4.
8. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nature Reviews Genetics*. 2011;12:671-682.
9. Sharma CM, Hoffmann S, Darfeuille F, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*. 2010;464:250-255.
10. Wang Y, Li X, Mao Y, Blaschek HP. Single-nucleotide resolution analysis of the transcriptome structure of *Clostridium beijerinckii* NCIMB 8052 using RNA-Seq. *BMC genomics*. 2011;12:479.
11. Giannoukos G, Ciulla DM, Huang K, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*. 2012;13:R23.
12. Li S, Dong X, Su Z. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling. *BMC genomics*. 2013;14:520.
13. Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. *Genome Research*. 2011;21:1498-1505.
14. ENCODE. The ENCODE Consortium: Standards, Guidelines and Best Practices for RNA-Seq.

- [https://genome.ucsc.edu/encode/protocols/dataStandards/ENCODE\\_RNaseq\\_Standards\\_V1.0.pdf](https://genome.ucsc.edu/encode/protocols/dataStandards/ENCODE_RNaseq_Standards_V1.0.pdf). 2011.
15. Landt S, Marinov G. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* 2012;22(9):1813-1831.
  16. Høvik H, Yu W-H, Olsen I, Chen T. Comprehensive transcriptome analysis of the periodontopathogenic bacterium *Porphyromonas gingivalis* W83. *Journal of bacteriology*. 2012;194:100-114.
  17. Mitschke J, Georg J, Scholz I, et al. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proceedings of the National Academy of Sciences*. 2011;108:2124-2129.
  18. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH. Structure and complexity of a bacterial transcriptome. *Journal of bacteriology*. 2009;191:3203-3211.
  19. Soutourina OA, Monot M, Boudry P, et al. Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. *PLoS genetics*. 2013;9:e1003493.
  20. He S, Wurtzel O, Singh K, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature methods*. 2010;7:807-812.
  21. Peano C, Pietrelli A, Consolandi C, et al. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp*. 2013;3:1-11.
  22. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics*. 2012;13:484.
  23. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*. 2014;15:121-132.
  24. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. *Nature methods*. 2013;10:325-327.
  25. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*. 2010;38:e131--e131.
  26. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-Sequencing data. *Bmc Bioinformatics*. 2011;12:290.
  27. Nejepinska J, Malik R, Moravec M, Svoboda P. Deep sequencing reveals complex spurious transcription from transiently transfected plasmids. *PloS one*. 2012;7:e43283.
  28. Raghavan R, Sloan DB, Ochman H. Antisense transcription is pervasive but rarely conserved in enteric bacteria. *mBio*. 2012;3(4):e00156-12.

29. O'Neil S, Emrich S. Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC genomics*. 2013;14:465.
30. Smith-Unna RD, Bournnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. *Genome Research* 2015;26(8):1134-1144.
31. Jones SW, Paredes CJ, Tracy B, et al. The transcriptional program underlying the physiology of Clostridial sporulation. *Genome Biol.* 2008;9:R114.
32. Paredes CJ, Alsaker KV, Papoutsakis ET. A comparative genomic view of Clostridial sporulation and physiology. *Nature Reviews Microbiology*. 2005;3:969-978.
33. Al-Hinai MA, Jones SW, Papoutsakis ET. The Clostridium Sporulation Programs: Diversity and Preservation of Endospore Differentiation. *Microbiology and Molecular Biology Reviews*. Mar 2015;79(1):19-37.
34. Nölling J, Breton G, Omelchenko MV, et al. Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. *Journal of bacteriology*. 2001;183:4823-4838.
35. Fischer RJ, Helms J, Dürre P. Cloning, sequencing, and molecular analysis of the sol operon of Clostridium acetobutylicum, a chromosomal locus involved in solventogenesis. *Journal of bacteriology*. 1993;175:6959-6969.
36. Gerischer U, Dürre P. Cloning, sequencing, and molecular analysis of the acetoacetate decarboxylase gene region from Clostridium acetobutylicum. *Journal of bacteriology*. 1990;172:6907-6918.
37. Gerischer U, Dürre P. mRNA analysis of the adc gene region of Clostridium acetobutylicum during the shift to solventogenesis. *Journal of bacteriology*. 1992;174:426-433.
38. Nair RV, Bennett GN, Papoutsakis ET. Molecular characterization of an aldehyde/alcohol dehydrogenase gene from Clostridium acetobutylicum ATCC 824. *Journal of bacteriology*. 1994;176:871-885.
39. Narberhaus F, Bahl H. Cloning, sequencing, and molecular analysis of the groESL operon of Clostridium acetobutylicum. *Journal of bacteriology*. 1992;174:3282-3289.
40. Venkataramanan KP, Min L, Hou SY, et al. Complex and extensive post-transcriptional regulation revealed by integrative proteomic and transcriptomic analysis of metabolite stress response in Clostridium acetobutylicum. *Biotechnology for Biofuels*. Jun 2015;8:81.
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114-2120.
42. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078-2079.

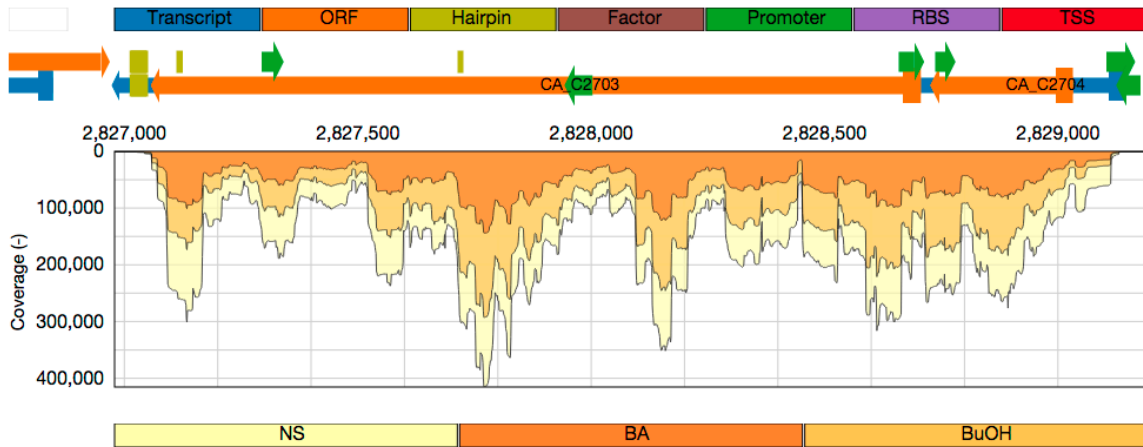
43. Paredes CJ, Rigoutsos I, Papoutsakis ET. Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic acids research*. 2004;32:1973-1981.
44. Bostock M, Ogievetsky V, Heer J. D3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*. 2011;17:2301-2309.
45. Ralston MT. *Assembling improved gene annotations in Clostridium acetobutylicum with RNA sequencing*. <http://udspace.udel.edu/handle/19716/16988> . Newark, DE, USA: Computer Science, University of Delaware; 2015.
46. Alsaker KV, Papoutsakis ET. Transcriptional program of early sporulation and stationary-phase events in *Clostridium acetobutylicum*. *Journal of Bacteriology*. Oct 2005;187(20):7103-7118.
47. Alsaker KV, Paredes C, Papoutsakis ET. Metabolite Stress and Tolerance in the Production of Biofuels and Chemicals: Gene-Expression-Based Systems Analysis of Butanol, Butyrate, and Acetate Stresses in the Anaerobe *Clostridium acetobutylicum*. *Biotechnology and Bioengineering*. Apr 2010;105(6):1131-1147.
48. Alsaker KV, Spitzer TR, Papoutsakis ET. Transcriptional analysis of *spo0A* overexpression in *Clostridium acetobutylicum* and its effect on the cell's response to butanol stress. *Journal of Bacteriology*. Apr 2004;186(7):1959-1971.
49. Tomas CA, Beamish J, Papoutsakis ET. Transcriptional analysis of butanol stress and tolerance in *Clostridium acetobutylicum*. *Journal of Bacteriology*. Apr 2004;186(7):2006-2018.
50. Tomas CA, Welker NE, Papoutsakis ET. Overexpression of *groESL* in *Clostridium acetobutylicum* results in increased solvent production and tolerance, prolonged metabolism, and changes in the cell's transcriptional program. *Applied and Environmental Microbiology*. Aug 2003;69(8):4951-4965.
51. Terracciano JS, Rapaport E, Kashket ER. Stress-and growth phase-associated proteins of *Clostridium acetobutylicum*. *Applied and environmental microbiology*. 1988;54:1989-1995.
52. Pich A, Narberhaus F, Bahl H. Induction of heat shock proteins during initiation of solvent formation in *Clostridium acetobutylicum*. *Applied Microbiology and Biotechnology*. 1990;33:697-704.
53. Narberhaus F, Giebler K, Bahl H. Molecular characterization of the *dnaK* gene region of *Clostridium acetobutylicum*, including *grpE*, *dnaJ*, and a new heat shock gene. *Journal of bacteriology*. 1992;174:3290-3299.
54. Wang Q, Venkataramanan KP, Huang H, Papoutsakis ET, Wu CH. Transcription factors and genetic circuits orchestrating the complex,

- multilayered response of *Clostridium acetobutylicum* to butanol and butyrate stress. *BMC systems biology*. 2013;7:120.
55. Cornillot E, Nair RV, Papoutsakis ET, Soucaille P. The genes for butanol and acetone formation in *Clostridium acetobutylicum* ATCC 824 reside on a large plasmid whose loss leads to degeneration of the strain. *Journal of Bacteriology*. Sep 1997;179(17):5442-5447.
  56. Fischer RJ, Helms J, Dürre P. Cloning, sequencing, and molecular analysis of the sol operon of *Clostridium acetobutylicum*, a chromosomal locus involved in solventogenesis. *J Bacteriol*. Nov 1993;175(21):6959-6969.
  57. Nair RV, Bennett GN, Papoutsakis ET. MOLECULAR CHARACTERIZATION OF AN ALDEHYDE/ALCOHOL DEHYDROGENASE GENE FROM CLOSTRIDIUM-ACETOBUTYLICUM ATCC-824. *Journal of Bacteriology*. Feb 1994;176(3):871-885.
  58. Nair RV, Papoutsakis ET. Expression of plasmid-encoded aad in *Clostridium acetobutylicum* M5 restores vigorous butanol production. *J Bacteriol*. Sep 1994;176(18):5843-5846.
  59. Wiesenborn DP, Rudolph FB, Papoutsakis ET. Coenzyme A transferase from *Clostridium acetobutylicum* ATCC 824 and its role in the uptake of acids. *Appl Environ Microbiol*. Feb 1989;55(2):323-329.
  60. Gerischer U, Dürre P. mRNA analysis of the adc gene region of *Clostridium acetobutylicum* during the shift to solventogenesis. *J Bacteriol*. Jan 1992;174(2):426-433.
  61. Petersen DJ, Cary JW, Vanderleyden J, Bennett GN. Sequence and arrangement of genes encoding enzymes of the acetone-production pathway of *Clostridium acetobutylicum* ATCC824. *Gene*. Jan 1993;123(1):93-97.
  62. Jones AJ, Fast AG, Clupper M, Papoutsakis ET. Small and low, but potent: the complex regulatory role of the small RNA SolB on solventogenesis in *Clostridium acetobutylicum*. *Applied and Environmental Microbiology*. May 4, 2018 2018;84(14):e00597-00518.  
<https://doi.org/00510.01128/AEM.00597-00518>.
  63. Harris LM, Blank L, Desai RP, Welker NE, Papoutsakis ET. Fermentation characterization and flux analysis of recombinant strains of *Clostridium acetobutylicum* with an inactivated solR gene. *Journal of Industrial Microbiology & Biotechnology*. Nov 2001;27(5):322-328.
  64. Nair RV, Green EM, Watson DE, Bennett GN, Papoutsakis ET. Regulation of the sol locus genes for butanol and acetone formation in *Clostridium acetobutylicum* ATCC 824 by a putative transcriptional repressor. *Journal of Bacteriology*. Jan 1999;181(1):319-330.
  65. Harris LM, Welker NE, Papoutsakis ET. Northern, morphological, and fermentation analysis of spo0A inactivation and overexpression in

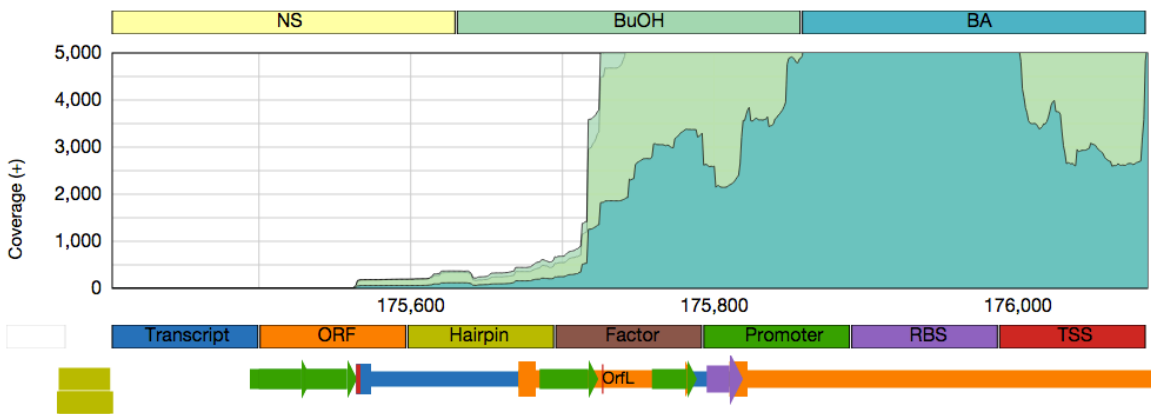
- Clostridium acetobutylicum ATCC 824. *Journal of Bacteriology*. Jul 2002;184(13):3586-3597.
66. Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods*. 2010;7:709-715.
  67. Piros SJ, Husseini GA. Preliminary Modeling of Transfer RNA Kinetics in the Cytoplasm of Escherichia coli Bacteria. *Advanced Science Letters*. 2010;3:28-36.
  68. Chen Y, Indurthi DC, Jones SW, Papoutsakis ET. Small RNAs in the genus Clostridium. *MBio*. 2011;2:e00340-00310.
  69. Fierromonti IP, Reid SJ, Woods DR. DIFFERENTIAL EXPRESSION OF A CLOSTRIDIUM-ACETOBUTYLICUM ANTISENSE RNA - IMPLICATIONS FOR REGULATION OF GLUTAMINE-SYNTHEASE. *Journal of Bacteriology*. Dec 1992;174(23):7642-7647.
  70. Cai X, Servinsky M, Kiel J, Sund C, Bennett GN. Analysis of redox responses during TNT transformation by Clostridium acetobutylicum ATCC 824 and mutants exhibiting altered metabolism. *Applied microbiology and biotechnology*. 2013;97:4651-4663.
  71. Bennett GN, Rudolph FB. THE CENTRAL METABOLIC PATHWAY FROM ACETYL-COA TO BUTYRYL-COA IN CLOSTRIDIUM-ACETOBUTYLICUM. *Fems Microbiology Reviews*. Oct 1995;17(3):241-249.
  72. Boynton ZL, Bennett GN, Rudolph FB. Cloning, sequencing, and expression of clustered genes encoding beta-hydroxybutyryl-coenzyme A (CoA) dehydrogenase, crotonase, and butyryl-CoA dehydrogenase from Clostridium acetobutylicum ATCC 824. *Journal of Bacteriology*. Jun 1996;178(11):3015-3024.



**Figures**

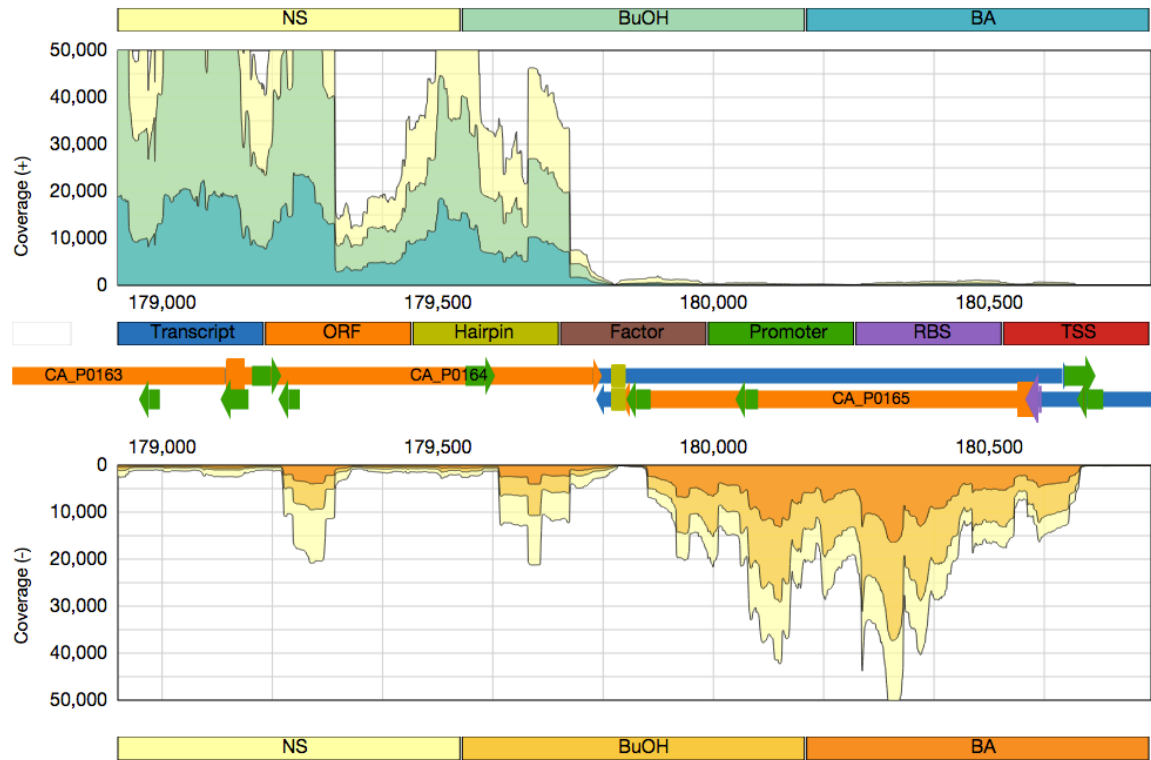


**Figure 1:** Genome browser visualization of the *groES/EL* loci. Information is labeled by color, with annotation and coverage vector legends at the top and bottom of the figure, respectively. The coverage trend is consistent with previously determined transcript boundaries and promoter and terminator motifs. RNAseq data collected from: NS, no stress conditions; BA, butyrate stress conditions; or BuOH, butanol stress conditions. These yellow, orange colors are used for transcripts from the negative DNA strand.

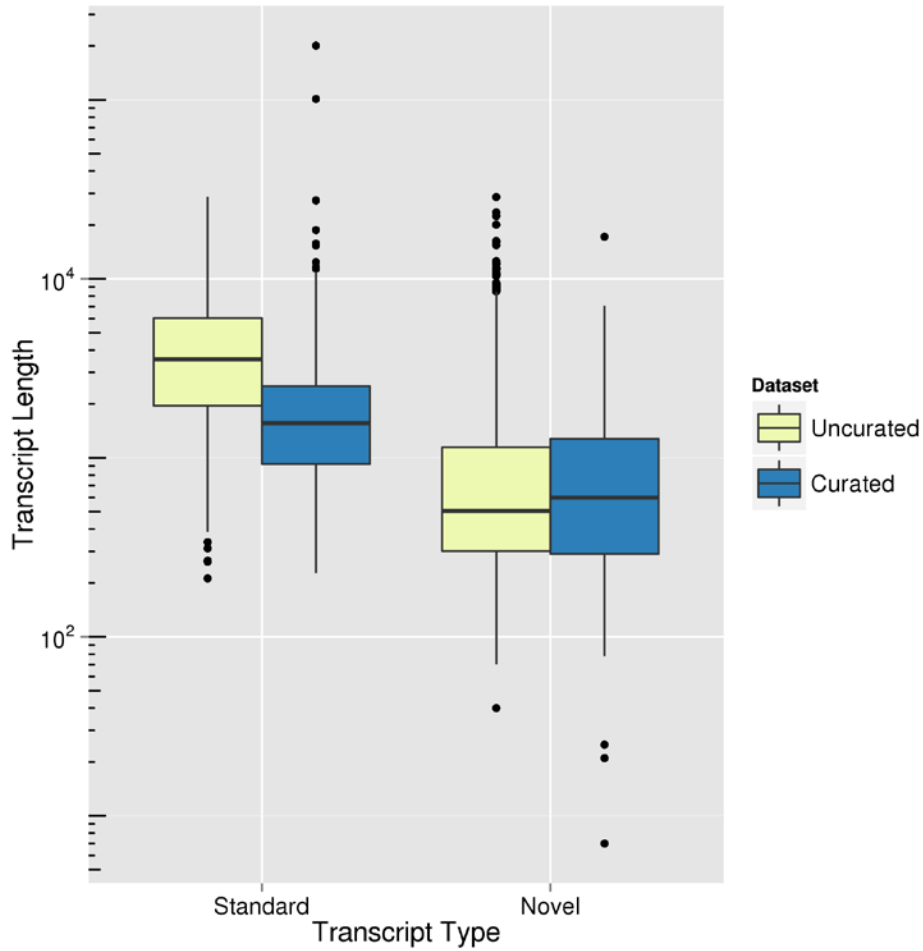


**Figure 2a:** Transcription start sites of the *sol* operon. The distal transcription start site location and expression level matches that of previous studies (red, left; see text for details). The location of a more abundant transcription start site (red, center) also matches

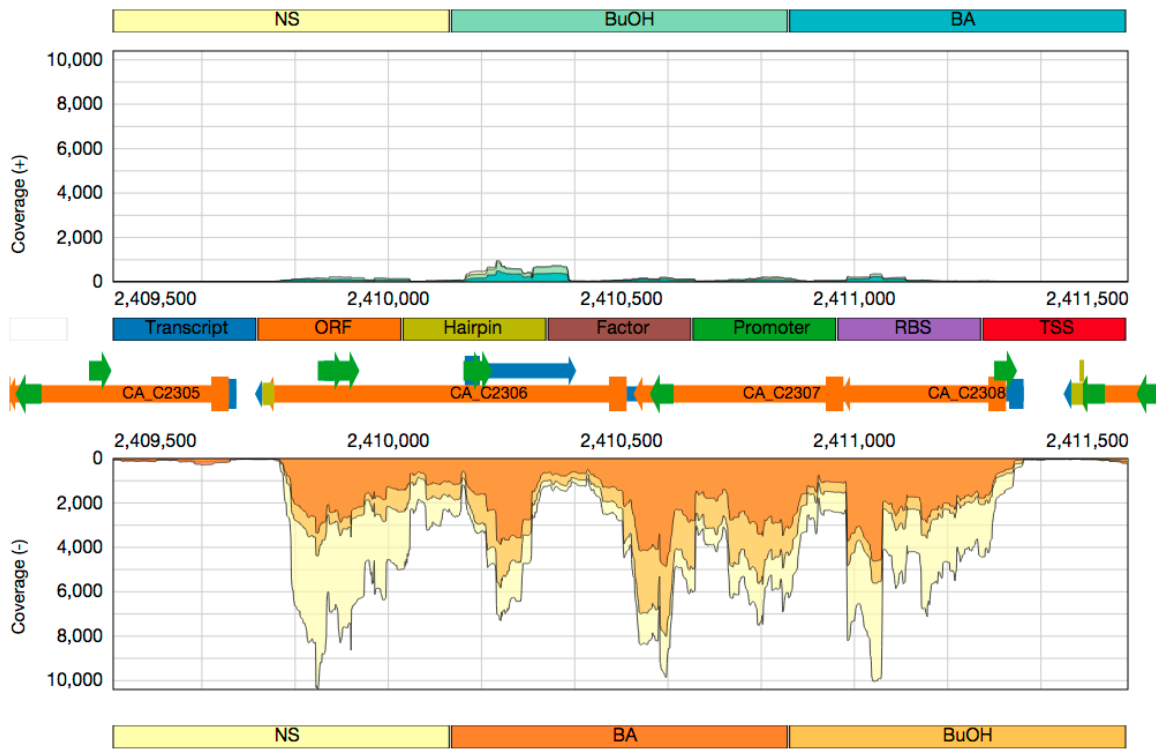
the proximal TSS from these studies. RNAseq data collected from: NS, no stress conditions; BA, butyrate stress conditions; or BuOH, butanol stress conditions. Light yellow, green colors are used for transcripts from the positive DNA strand.



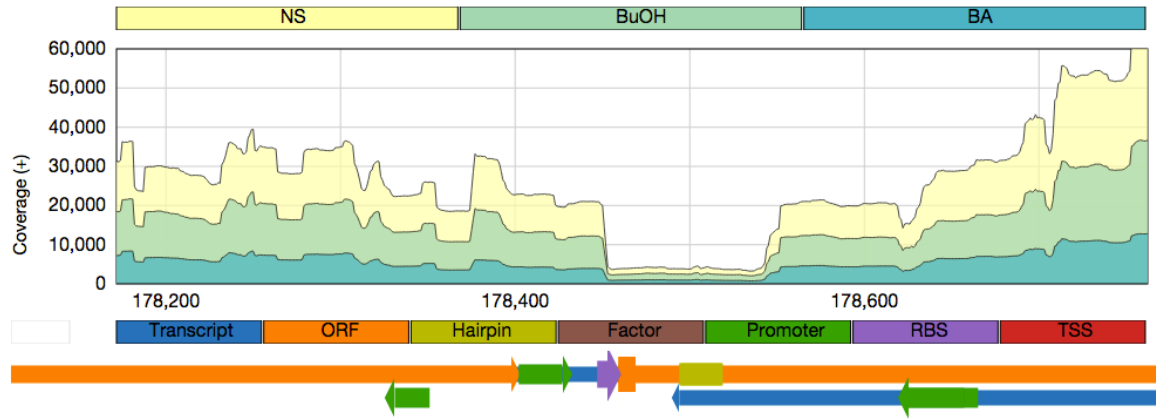
**Figure 2b:** Genome browser visualization of the *sol* locus, with annotations labeled as above. A bifunctional Rho-independent terminator (yellow, center) stops transcription of the *sol* operon (upper track) and the *adc* gene (lower track). Residual antisense signal from the *adc* gene was sufficiently complex to trigger misassembly of the *sol* transcript, shown by the extended assembled transcript on the forward strand of the central annotation track (blue from center to far right). See legends of Fig. 1 and 2a for additional explanations.



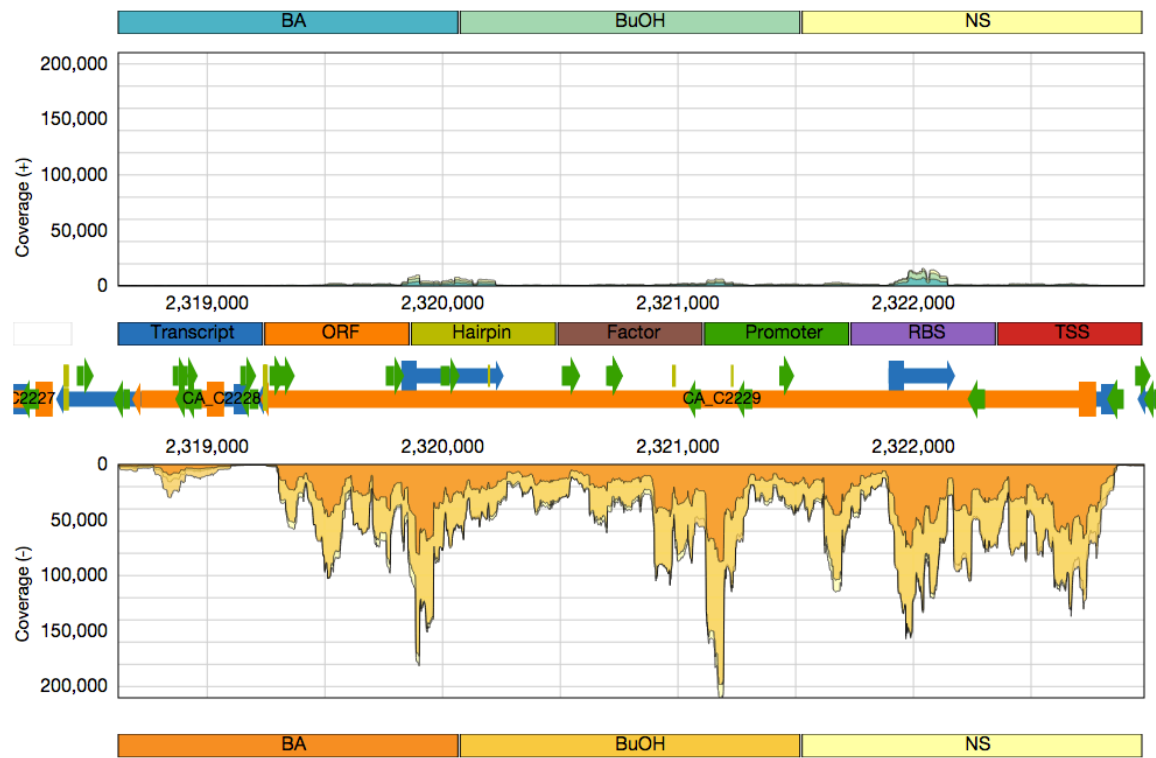
**Figure 3:** Curation of the standard and novel transcripts improved the consistency of the distributions with knowledge of transcript lengths in the model prokaryote *Escherichia coli*.



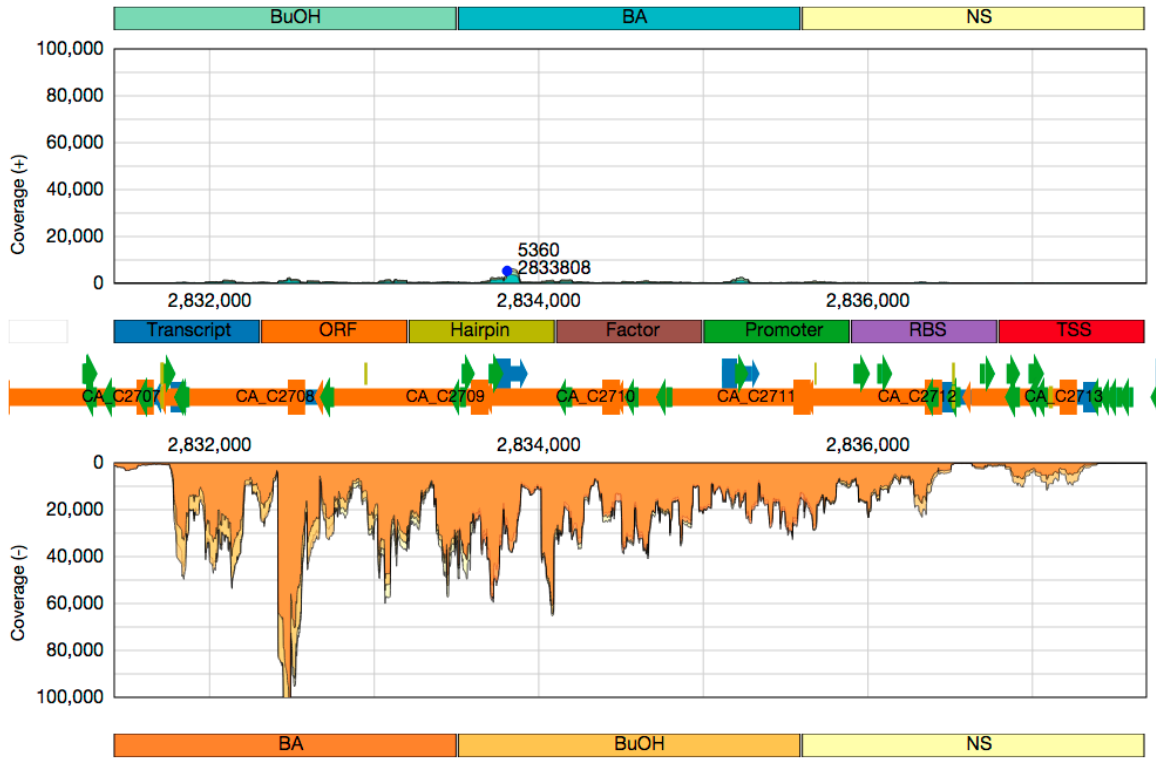
**Figure 4:** Sigma factor F (*sigF*) operon is shown on the bottom track (negative DNA strand). Antisense signal is observed on the forward (positive) DNA strand, displaying different stress responsive behavior. See legends of Figures 1 and 2 for additional explanations.



**Figure 5:** The *adhe1-ctfA* IGR (intergenic region) contains a promoter and ribosome binding site (RBS) for a second putative transcript that includes only *ctfA* and *ctfB*. See legend of Figure 2 for additional explanations.



**Figure 6:** The *hydA* transcript with antisense transcription on the forward strand. See legends of Figures 1 and 2 for additional explanations.



**Figure 7:** The BCS operon region displays antisense transcription on the forward (positive) strand.