1 LacI Transcriptional Regulatory Networks in *Clostridium thermocellum* DSM1313

2

3 Charlotte M. Wilson[a,b,] , Dawn M. Klingeman [a,b], Caleb Schlachter [a,b], Mustafa H. Syed [a,b], Chia-
4 wei Wu [a,b], Adam M. Guss [a,b], and Steven D. Brown [a,b,#].

5

6 BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831,
7 USA[a]; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831,
8 USA[b];

9

10 Running Head: *C. thermocellum* LacI Regulons

11

12 #Address correspondence to Steven D. Brown, brownsd@ornl.gov.

13

14 This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-
15 00OR22725 with the U.S. Department of Energy.

16

17

Abstract

Organisms regulate gene expression in response to the environment to coordinate metabolic reactions. *Clostridium thermocellum* expresses enzymes for both lignocellulose solubilization and its fermentation to produce ethanol. One LacI regulator termed GlyR3 in *C. thermocellum* ATCC27405 was previously identified as a repressor of neighboring genes with repression relieved by laminaribiose (a β-1,3 disaccharide). To better understand the three *C. thermocellum* LacI regulons, deletion mutants were constructed using the genetically tractable DSM1313 strain. DSM1313 *lacI* genes Clo1313_2023, Clo1313_0089, and Clo1313_0396 encode homologs of GlyR1, GlyR2 and GlyR3 from strain ATCC27405, respectively. Growth on cellobiose or pretreated switchgrass was unaffected by any of the gene deletions under pH controlled fermentations. Global gene expression patterns from time course analyses identified glycoside hydrolase genes encoding hemicellulases, including cellulosomal enzymes, that were highly up-regulated (5 to 100 fold) in the absence of each LacI regulator, suggesting these were repressed under wild type conditions and that relatively few genes were controlled by each regulator under the conditions tested. Clo1313_2022, encoding lichenase enzyme LicB, was derepressed in a Δ*glyR1* strain. Higher Clo1313_1398 expression, which encodes the Man5A mannanase, was observed in a Δ*glyR2* strain and α-mannobiose was identified as a probable inducer for GlyR2 regulated genes. For the Δ*glyR3* strain, up regulation of the two genes adjacent to *glyR3* in the *celC-glyR3-licA* operon was consistent with earlier studies. Electrophoretic mobility shift assays have confirmed LacI transcription factor binding to specific regions of gene promoters.

(250 words max) now at 237

Applied and Environmental Microbiology

47 **Importance (150 words max, now at 87)**

48 Understanding *C. thermocellum* gene regulation is of importance for improved fundamental
49 knowledge of this industrially relevant bacterium. Most LacI transcription factors regulate local
50 genomic regions; however, a small number of those genes encode global regulatory proteins with
51 extensive regulons. This study indicates there are small specific small *C. thermocellum* LacI
52 regulons. The identification of LacI repressor activity for hemicellulase gene expression is a key
53 result of this work and will add to the small body of existing literature on the area of gene
54 regulation in *C. thermocellum*.

55 **Introduction**

56 *Clostridium thermocellum* is an anaerobic thermophile with an elaborate and highly efficient
57 extracellular enzyme system for solubilizing lignocellulosic biomass. *C. thermocellum* is of
58 biotechnological interest due to its potential for lignocellulose conversion to fuels.
59 *C. thermocellum* produces cellulosomes, which are multiprotein structures consisting of a protein
60 platform (scaffoldins) with attached cellulolytic enzymes. These enzymes are bound to
61 scaffoldins via Type 1 cohesion-dockerin domain interactions and this large complex is anchored
62 to the cell wall by secondary scaffoldins and Type 2 cohesin-dockerin interations (1). In addition,
63 the bacterium produces free enzymes and "cell-free" cellulosomes (2) and it is a candidate for
64 cellulosic ethanol production (3). Wild-type *C. thermocellum* can utilize hexose sugars but not
65 pentose sugars as fermentation substrates, despite having eight genes encoding xylanases (1).
66 The organism has characterized ABC transporters for the uptake of cellodextrins ranging from
67 G2-G5 with one, CbpA specific for cellotriose and another specific for laminaribiose (4).
68 Understanding *C. thermocellum* gene regulation is of importance for improved fundamental
69 knowledge of this industrially relevant bacterium.
70
71 *C. thermocellum* regulates gene expression in response to carbon sources and growth rates (5-7).
72 The expression of the cellulosomal proteins is influenced by substrate availability and
73 extracellular polysaccharides (8). Most cellulase and hemicellulase genes of *C. thermocellum* are
74 scattered throughout the genome in small clusters, in contrast to the larger operons in mesophilic
75 cellulolytic clostridia (9-16). Methods of sensing and coordinating gene expression must
76 therefore exist to allow the organism to utilize cellodextrins in plant biomass (17). Altered gene
77 expression in response to substrate availability is expected to affect the composition of the

78 cellulosome in terms of binding available dockerin containing enzymes to the cohesion domains
79 on the scaffoldin (CipA) backbone (6). *C. thermocellum* DSM1313 was used in this study due to
80 the existence of a genetic system and the absence of an equivalent system in the type strain
81 *C. thermocellum* ATCC 27405. The genomes for strains DSM1313 and ATCC 27405 are very
82 similar with average nucleotide identities (ANIs) of 99.6 and 99.3 % in reciprocal genome
83 comparisons (18). Strain DSM1313 encodes more than 100 transcription regulators. However,
84 only a small proportion of *C. thermocellum* transcriptional regulators have been studied.
85 Previous work has demonstrated the inducible expression of the *celC-glyR3-licA* operon (11, 19),
86 the involvement of alternative sigma factors with their cognate anti-sigma factors in the
87 regulation of cellulolytic genes including *celS* (17, 20, 21) and histidine kinase genes involved in
88 the regulation of sporulation (22).
89
90 *C. thermocellum* strains DSM1313 and ATCC 27405 each encode three LacI transcriptional
91 regulators (GlyR1-3), the homologs of these regulators in the two strains are 100% identical in
92 respective comparisons. Within *C. thermocellum* strain comparisons between LacI proteins give
93 similarity values of ~24-42% identity at amino acid level. LacI family members can sense sugar
94 and non-carbohydrate effectors via a C-terminal effector binding domain, with the majority
95 regulating carbohydrate utilization genes, and have a characteristic N-terminal helix-turn-helix
96 DNA binding domain (23). The number LacI regulators varies per genome, with most having
97 fewer than ten such regulators and some Actinobacteria having up to 32. The characterized
98 ATCC 27405 genes *celC-glyR3-licA* (Cthe_2807, 2808, 2809) are homologous to DSM1313
99 genes Clo1313_0395, 0396, 0397. CelC and LicA, a glycoside hydrolase 5 family member and
100 glucan endo-1,3-beta-D-glucosidase, respectively, that are not cellulosome associated and are

101 instead members of the *C. thermocellum* free cellulolytic enzyme system (1). The activity of the
102 GlyR3 regulator on this three gene operon was demonstrated to have repressor activity, which
103 was relieved in the presence of laminaribiose, a β-1,3-linked glucose dimer. GlyR3 was shown to
104 bind to an 18 bp near-palindromic sequence upstream of the start codon of Cthe_2807 (19).
105
106 To date, the majority of *C. thermocellum* genetic engineering studies have focused on improving
107 fuel production (24-30), but little genetic work has targeted understanding regulatory pathways.
108 Here, we combine use of gene deletions with transcriptomics and DNA binding assays to gain
109 insights into the regulatory networks of the three LacI transcription factors encoded in the
110 genome of *C. thermocellum* DSM1313.
111
112 **Materials and Methods**
113 **Bacterial strains and mutant construction**
114 Plasmids were constructed with standard methods (31), including yeast gap repair (32) and
115 Gibson Assembly (33) using the Gibson Assembly Cloning Kit (New England Biolabs, Ipswich,
116 MA, USA), resulting in plasmids pAMG488 for deletion of Clo1313_0396, pCS2 for deletion of
117 Clo1313_0089, and pCS6 for deletion of Clo1313_2023. Plasmids were isolated from
118 *Escherichia coli* strain BL21, such that the DNA was not methylated by Dcm methylase in
119 *E. coli* prior to transformation into *C. thermocellum* (34). *C. thermocellum* transformations and
120 gene deletions were performed as previously described (25, 35) using primers shown for
121 construction and strain verification (Table 1). Annotated plasmid maps are provided (Data Set
122 S1-S3 in the supplemental material). Briefly, plasmid DNA was electroporated into
123 *C. thermocellum Δhpt*, and plated on solid CTFUD medium containing 5 μg ml^{-1} thiamphenicol

(TM; Sigma-Aldrich, St. Louis, MO) to select for transformation. Colonies were picked into liquid CTFUD medium supplemented with TM, followed by plating on solid CTFUD medium containing TM and 50 μg ml^{-1} 5-fluoro-2'-deoxyuridine (Sigma-Aldrich, St. Louis, MO) to select for merodiploids strains. After single colony purification, colonies were picked into liquid CTFUD medium containing TM. Strains were subcultured in CTFUD medium without antibiotics, followed by plating on solid CTFUD medium supplemented with 500 μg ml^{-1} 8-azahypoxanthine (Acros Organics, Pittsburgh, PA). Strains were again single colony purified and picked into liquid CTFUD medium, and the resulting unmarked gene deletions were verified by PCR.

Fermentation growth conditions

Bioreactors were prepared for *C. thermocellum* culture by autoclaving cellobiose, water and resazurin (total volume 800 ml) in 1-liter Biostat Q-Plus bioreactors (Sartorius Stedim, Gottingen, Germany) for 30 min. Reactors were sparged with nitrogen gas for one hour and a sterile cocktail of the remaining MTC media components prepared (36). A cocktail (100 ml) was added to each reactor and sparging with nitrogen continued overnight with observable loss of color from resazurin reduced. The initial vessel pH was checked post-autoclaving using an external pH probe and slight adjustments were made to the pH calibration controlling the reactors as necessary. The fermenters were run at 58°C, 100 rpm and pH controlled at 7.0 with 3N NaOH. Immediately prior to inoculation the gas was switched from 100% Nitrogen to a 80% Nitrogen;20% CO2 mix and the pH of the reactor allowed to drop to ~pH 6.2. The automatic base addition pump was switched on and the pH of the reactor brought back up to pH 7.0. The

147 bioreactors were subsequently inoculated using overnight inoculum cultures of *C. thermocellum*
148 DSM1313 strains that had been grown anaerobically in 500 mL bottles of MTC containing 5g/L
149 cellobiose. Inoculum aliquots of 100 mL were subcultured into each reactor (total volume 1 L)
150 for a final inoculum of ~10%. Post-inoculation the gas was switched from passing through the
151 growth medium to just passing through the head space of the reactor and was only used to create
152 positive vessel pressure during sampling, otherwise the gas sparge remained off during bacterial
153 growth. Triplicate fermentations were performed for each stain. Growth was tracked by base
154 addition and Optical Density (OD_{600}) measurements taken every two hours. Samples (50 mL)
155 were removed at mid-log phase (OD_{600} of 0.7-0.95), late log phase (OD_{600} of 1.8-2.0) and in
156 stationary phase approximately 30 min after cultures stopped producing acid (OD_{600} 2.0-2.2).
157
158 **Genomic DNA preparation and RNA extraction**
159 Genomic DNA for each of the *lacI* deletion strains was prepared using a Wizard Genomic DNA
160 Purification Kit (Promega Corp., Madison, WI, USA). For total RNA preparation, cells pelleted
161 from a 50 mL sample drawn from each fermenter were resuspended in 3 mL of TRIzol reagent
162 (Invitrogen, Carlsbad, CA, USA). A 1 ml aliquot of the TRIzol cell suspension was used for cell
163 lysis by bead beating with 0.8 g of 0.1mm glass beads (BioSpec Products, Bartlesville, OK,
164 USA) and 3 X 20 s bead beating treatments at 6,500 rpm in a Precellys 24 high-throughput tissue
165 homogenizer (Bertin Technologies, Montigny-le-Bretonneux, France). The RNA from each cell
166 lysate was purified, DNaseI treated, quantity and quality assessed as previously described (7).
167 Purified RNA was depleted of rRNA using the Ribo-Zero rRNA Removal kit for Gram Positive
168 Bacteria (Epicentre, Madison, WI, USA). The rRNA depleted RNA sample was concentrated

169 with RNA Clean & Concentrate-5 (Zymo Research, Irvine, CA, USA) following the
170 manufacturer's protocol.
171
172 **Library preparation and DNA sequence data generation**
173 For genome resequencing, libraries were prepared using a TruSeq LT kit that employed a gel-
174 based size selection and subsequently sequenced on an MiSeq instrument (Illumina, San Diego,
175 CA, USA) using version 2 chemistry and 2 x 250 bp paired-end configuration. Depleted RNA
176 was used as the starting material for the ScriptSeq™ v2 RNA-Seq Library Preparation Kit
177 (Epicentre, Illumina compatible) utilising Fail Safe PCR Enzyme mix (Epicentre) and following
178 the manufacturer's protocol. cDNA tagged with Script-seq adaptors (1-12) was eluted with 20 µl
179 of Buffer EB provided in the Min-Elute PCR purification kit (Qiagen, Valencia, CA) according
180 to the Script-Seq™ mRNA-Seq Library preparation kit protocol. Twelve PCR cycles were used
181 during library amplification and samples were purified using Agencourt AMPure XP beads
182 (Beckman Coulter) according the library protocol and eluted with 20 µl of nuclease free water.
183 The final mRNA Seq library was quantified with a Qubit Fluorometer (Invitrogen) and library
184 quality was assessed with Agilent Bioanalyzer DNA 7500 Chip for DNA (Agilent, Santa Clara,
185 CA, USA). Samples were pooled into three groups to make a stock of 20 nM in 72 µl. A 40 µl
186 aliquot of the three pooled libraries was sequenced using a SR50 sequencing run on an Illumina
187 HiSeq 2500 platform (HudsonAlpha Genomic Services Laboratory; Huntsville, AL).
188
189 **Resequencing and RNA-seq analysis**
190 The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used for resequencing data
191 quality control. Illumina reads were trimmed if the median quality of reads at 5` or 3` ends was

192 below 30, and any reads with more than 10% of positions below a quality score of 30 were
193 filtered out. Bowtie2 (37) was used to align reads to the *C. thermocellum* DSM1313 genome
194 sequence (GenBank accession number CP002416.1) with default parameters, which generated
195 sam files, and subsequently samtools (38) was used to call variants (with minimum read depth of
196 -d 20 being specified). Putative SNPs were verified by PCR and Sanger sequencing analysis
197 using described primers (Table 1). Raw resequencing data are available from the NCBI Sequence
198 Read Archive (SRA)under accession SRP074474. For RNA-seq, reads were mapped to the
199 *C. thermocellum* DSM1313 genome using CLC Genomics Workbench version 8 (CLCbio) using
200 the default settings for prokaryote genomes. Counts of uniquely mapped reads were analyzed for
201 differential gene expression by DESeq2 (39). Filtering was applied to identify those genes with
202 an FDR <0.05 and a greater than a $\log_2$ of +/-1 for differential gene expression.
203
204 **Construction of recombinant expression plasmids**
205 Each *lacI* gene was cloned into plasmid pTXB1 (New England Biolab (NEB), Ipswich,
206 Massachusetts, USA) using primers indicated (Table 1). The native stop codon was left off the
207 construct so a fusion chitin binding domain tag would be expressed on cloning into pTXB1.
208 These were transformed into 5-α High Efficiency Competent *E. coli* (NEB) and constructs
209 verified by PCR and Sanger sequencing. Plasmids were isolated and transformed into *E. coli*
210 strain T7 Express High Efficiency Competent *E. coli* (NEB) for expression.
211
212 **Protein purification**
213 A single colony from an LB plate was used to inoculate 10 ml of LB medium with ampicillin
214 (100µg/ml). The culture was grown overnight at 37°C with 200 rpm shaking. A 2 ml aliquot of

215 the overnight culture was used to inoculate a 500 mL volume of LB medium and grown at 37°C
216 with 200 rpm shaking until an OD_{600} of ~0.6 was reached. Protein expression was induced with
217 filter sterilized IPTG (final 0.5mM) and incubation followed overnight at 15°C with 200 rpm
218 shaking. Cells were harvested by centrifugation in two lots of 250 ml culture at 8000 x *g* for 15
219 min at 4°C. Cell pellets were resuspended in 25 mL of column buffer (20 mM Tris-HCl pH 8.5,
220 0.5 M NaCl, 1 mM EDTA, 0.1% Tween-20) and transferred to 50 mL tubes. Cells were lysed by
221 sonication on ice using a microtip, with 60 cycles of 15 s on, 30 s off with the power set at 4
222 (Misonix Sonicator 4000, Misonix Inc, Farmingdale, NY, USA) Cell debris was removed by
223 centrifugation at 7919 x *g* for 30 min at 4°C. The fusion protein was purified using chitin resin
224 (NEB) following the manufacturer's protocols. After binding the fusion protein to the column
225 and washing of the bound material, the LacI protein was cleaved from the CBD fusion protein by
226 incubating the column with column buffer containing 50 mM DTT (~20 h at room temperature).
227 Eluted fractions (1.5 ml) were analyzed by SDS-PAGE and fractions observed to contain cleaved
228 protein were pooled and the column buffer exchanged for PBS using Amicon Ultra 0.5ml 10kDa
229 spin columns (EMD Millipore, Billerica, MA, USA). Protein concentration was determined by
230 Bradford Assay using BSA standards.
231
232 **Electrophoretic mobility shift assay**
233 In vitro binding of each LacI transcription factor to candidate promoters of genes showing
234 differential patterns of gene expression was assessed by EMSA using fluorescence based
235 detection EMSA kit (Life Technologies, Carlsbad, CA, USA). Target DNA fragments were
236 chosen based on the results of the RNA-seq experiment with candidate genes showing strong up-
237 regulation in the absence of a particular LacI transcription factor. Genomic regions upstream of

candidate responsive genes were amplified by PCR (see Table 1 for primers) and purified using a PCR Purification kit (Qiagen). Increasing amounts of each purified LacI transcription factor (0089: 0.016-1.5µM, 0396: 0.06-2.5 µM, 2023: 0.016-1.5µM final concentration) were incubated with purified target DNA (0.02-0.03 µM) for 30 min at room temperature using 1x EMSA binding buffer provided in the kit and 10 µL reactions. After the 30 min incubation, 2 µL of DNA loading dye was added to each tube and the samples were separated on a 6% TBE non-denaturing polyacrylamide gel (Life Technologies) at 125 V for 60 min in pre-chilled 0.5X TBE buffer. Gels were rinsed with DI water and stained with SYBR Green for 30 min in the dark at room temperature on a shaking platform set to 50 rpm. The gels were rinsed with ~150 ml deionized water twice and visualized using a transilluminator GelDoc (UVP) with a UV wavelength of 302nm and a SYBR filter.

Sugar disruption assays

The ability of various sugars (mannobiose, cellobiose, laminaribiose, maltose, sucrose, glucose lactose, galactose, arabinose, xylose) to disrupt LacI proteins binding to DNA fragments were tested in EMSA assays. EMSA tests were conducted using a Lightshift EMSA kit (Thermo Fisher Scientific, Waltham, MA USA). Probes were made by using biotin labeled primers (IDT) to create labeled probes (Table 1). Purified transcription factors were the same as those used in the fluorescent EMSA assays. EMSA experiments were performed according to the manufacturer's instructions with each reaction containing 1x binding buffer, 2.5% (v/v) glycerol, 5mM $MgCl_2$, 50 ng.μL^{-1} poly (dI-dC), and 0.05% (v/v) NP-40. Each reaction included 20 fmol of biotin labeled probe and 15 nM of purified transcription factor. Control competitive binding reactions included 4 pmol of unlabelled probe. Sugars were included in experiments at indicated

quantities. Assays were incubated at room temperature for 30 min and analyzed on 4% polyacrylamide gels in 0.5% TBE buffer. EMSA gels were electroblotted onto biodyne membrane and signal developed using Chemiluminescent Nucleic Acid Detection Module kit (Thermo Fisher Scientific).

Data Availability

Raw RNA-Seq data have been deposited in NCBI SRA under accession SRP057818 and gene expression data under NCBI GEO accession GSE68423.

Results

Deletion of the LacI regulators

Three genes annotated in the *C. thermocellum* DSM1313 genome as LacI transcriptional regulators were deleted as previously described (35). Deletion of each locus was confirmed by PCR and Sanger sequencing analysis. In addition, genotypes were analyzed by resequencing analysis by mapping respective Illumina reads to the *C. thermocellum* DSM1313 reference genome. In total, 112 differences were identified by resequencing the three deletion strains compared to the wild type reference genome (Table S1 in the supplemental material). Many SNPs were present in multiple strains with sixty-seven distinct mutations occurring across the three deletion strains, some of which may have also been present in the parental strain. Sixteen of these uniquely identified mutations were predicted to result in non-synonymous changes to coding regions and validated by Sanger sequencing of the region. Three of these sixteen mutations were mis-called during Illumina sequencing analysis and were the same as the DSM1313 reference genome. Six non-synonymous mutations occurred in homopolymer repeat

284 regions and were the same as the ATCC 27405 genome and are possibly errors in the DSM1313
285 reference genome. The remaining seven mutations were confirmed as present in at least one of
286 the strains.
287
288 Mutations present in the coding regions of each of the strains were as follows:
289 GlyR1/Clo1313_2023 had a SNP in Clo1313_0908 encoding adenine phosphoribosyltransferase,
290 a SNP in Clo1313_2711 encoding a hypothetical protein and a transposon insertion in
291 Clo1313_1891 encoding a protein with a predicted uridine kinase function (Fig. S1 in the
292 supplemental material). GlyR2/Clo1313_0089 had mutations that included a transposon insertion
293 into Clo1313_0145 encoding an aminotransferase (Fig. S1 in the supplemental material), a SNP
294 in Clo1313_0908 encoding adenine phosphoribosyltransferase, which was distinct from the
295 mutation in GlyR1 and a SNP in Clo1313_0638 encoding a protein with a histone family DNA-
296 binding domain. Clo1313_0396 had a confirmed SNP in uracil phosphoribosyltransferase
297 (Clo1313_0185) and the same SNP in Clo1313_0908 as strain GlyR2. Mutations were not found
298 in genes directly related to sugar metabolism and fermentation.
299
300 **Growth and harvesting for transcriptomic analysis**
301 Little growth difference on the soluble cellobiose substrate was detected between the three *lacI*
302 deletion strains and the parental strain (Fig. 1) and the deletion strains had no apparent phenotype
303 differences when pretreated switchgrass was provided as the carbon source (Fig. S2 in the
304 supplemental material). For RNA-seq analyses, triplicate one liter bioreactors were grown for
305 each strain to monitor growth and harvest samples. Each strain reached a similar turbidity before
306 transitioning to stationary phase. Samples were collected from each bioreactor at mid

307 exponential phase (OD_{600} 0.7-1.0), late exponential phase (OD_{600} 1.8-2.0) and early stationary
308 phase (OD_{600} 2.0-2.2) to extract RNA for transcriptome profiling by RNA-sequencing. Reads per
309 samples ranged from 15.5 M to 36.5 M with an average of 25M reads across the 36 samples
310 (Table S2 in the supplemental material). Between 97.2% and 98.9% of the reads per sample were
311 uniquely mapped to the DSM1313 reference sequence with genome coverages ranging from 194-
312 460x (Table S2 in the supplemental material). Mapped reads were imported into the DESeq2
313 workflow for normalization and determination of differential gene expression between each
314 strain at a given time point and the parental strain at the same growth stage. Genes were
315 considered differentially expressed with a 2-fold difference in gene expression between a mutant
316 and wild type strain at a given growth stage and an FDR (Benjami-Hochberg method) adjusted
317 *p*-value of less than 0.05 (Table S3 in the supplemental material).
318
319 **LacI regulons**
320 A distribution analysis of mean read unique read counts for the *Δhpt* strain at each sampling
321 point revealed counts for the three *lacI* genes were in the upper 50th percentile under all growth
322 conditions. The *Δhpt* strain is the parental strain used in the production of the lacI deletion
323 mutants. Deletion of the *hpt* gene in *C. thermocellum* is necessary to create markerless deletion
324 strains. At mid exponential, late exponential and early stationary phase for; *glyR1* there were an
325 average of 3042, 2956, and 3864 reads respectively; *glyR2* 4651, 4648, and 2232 reads
326 respectively, and *glyR32* 1362, 1413, and 3864 reads respectively. In each deletion strain, a
327 dramatic up regulation of at least one gene occurred in the absence of a particular regulator. This
328 was not dependent on growth stage as levels of expression remained consistently up regulated by
329 greater than 5-fold across the growth stage transcriptome profiling. Numbers of differentially

330 expressed genes varied between strains and the growth stages (Table S3 in the supplemental
331 material). The focus of this study was derepressed genes across all growth conditions.
332
333 **GlyR1**
334 Deletion of the *glyR1* gene (Clo1313_2023) resulted in 13, 23 and 1092 differentially expressed
335 genes at mid exponential, late exponential and early stationary phase respectively (see Table 2
336 for summary and Table S3 in the supplemental material for complete details). Substantial up
337 regulation (>20-fold) of Clo1313_2022 occurred across all three time points in the absence of
338 *glyR1* (Table 3). Clo1313_2022 is divergently transcribed to *glyR1* and encodes LicB, a
339 lichenase enzyme with a Dockerin I domain, indicating this enzyme can bind to the cellulosome
340 scaffoldins. Clo1313_2024, downstream and on the opposite strand to *glyR1,* did not show
341 differential expression. The only other gene (Clo1313_2865) upregulated at mid exponential
342 growth phase encoded a protein of unknown function (DUF214). Down regulated across all three
343 time points sampled was an operon Clo1313_1886-1891 on the reverse strand. Mapped RNAseq
344 data suggests these genes are transcribed as a single unit from the complement strand as reads
345 were mapped across the intergenic regions of these six genes. Strain resequencing revealed an
346 IS3 transposon insertion near the start of Clo1313_1891 (Fig. S1 in the supplemental material),
347 which likely disrupted transcription of this operon in the *glyR1* deletion strain. The
348 Clo1313_1886-1891 genes were the most down regulated genes across all three time points, with
349 at least 40-fold less expression occurring in the stationary phase relative to the parental strain.
350 The Clo1313_1886-1891 genes are poorly characterized unfortunately, with predicted functions
351 including a Fe-S cluster domain, a putative anti-sigma regulatory factor. If the anti-sigma
352 regulatory factor annotation is correct, this gene could affect the expression of other

353 *C. thermocellum* genes and be contributing to the large set of differentially expressed genes in
354 the Stationary phase comparison with the parental strain.
355
356 **GlyR2**
357 Deletion of the *glyR2* gene (Clo1313_0089) resulted in 41, 26 and 46 genes significantly
358 differentially expressed at mid exponential, late exponential and early stationary phase
359 respectively (see Table 2 for summary and Table S3 in the supplemental material for complete
360 details). Clo1313_0090 encodes an S-layer domain-containing protein and is downstream and
361 transcribed divergently to *glyR2*. The Clo1313_0090 gene showed approximately two-fold
362 higher expression in the *glyR2* mutant strain compared to the parent strains only in stationary
363 phase and its conditional derepression indicates other regulation. Clo1313_0088 encodes a
364 hypothetical protein, is on the same strand and is immediately downstream of *glyR2* and it did
365 not show differential expression under the conditions tested (Fig. S3 and Table S3 in the
366 supplemental material). In strain *ΔglyR2*, up regulation of Clo1313_1398 (Cthe_0821) was
367 observed across all three sampling time points (Fig. 3). Clo1313_1398 encodes the Man5A
368 protein, which contains a Dockerin I domain required for attachment to the cellulosome
369 scaffoldin proteins. Man5A also has GH5 and CBM32 domains and has been demonstrated to
370 have a preference for mannooligosaccharides. During mid exponential phase, deletion of the
371 *glyR2* gene resulted in an up regulation of a further nine genes with CAZyme (Carbohydrate-
372 Active Enzymes) related functions plus a gene encoding a glycogen/starch synthase. The
373 expression patterns of three of these genes was maintained at a higher level in the *glyR2* deletion
374 strain compared to the parental during late log phase; however, by stationary phase these

differences between the two strains had diminished, aside from the dramatic up regulation of Clo1313_1398.

Eight genes with predicted functions in the pentose phosphate pathway and central carbon metabolism, Clo1313_0073-0080, encoding two transketolase domain proteins, phosphoglycerate mutase, and a glycerol kinase were down regulated across all three time points sampled for both *glyR2* and *glyR3* deletion strains compared to the parental strain (Table S3 in the supplemental material). One gene, Clo1313_0638, annotated as encoding a histone family protein DNA-binding protein was down-regulated by more than 4-fold; however, the strain resequencing had identified a SNP present in the coding region (GI: 724743). This SNP led to an isoleucine to threonine change at the amino acid level. The *glyR3* deletion strain had the same mutation and also had a similar down regulation of the gene relative to the parental strain (see below), suggesting this pattern of gene expression was related to the SNP and not the respective *lacI* gene deletions.

GlyR3

Deletion of the previously described *C. thermocellum* LacI transcriptional regulator gene, *glyR3* (Cthe_2808/Clo1313_0396), resulted in 31, 65 and 75 genes considered significantly differentially expressed at mid exponential, late exponential and early stationary phase respectively (Table 1). Up regulation of the two genes adjacent to *glyR3* in the *celC-glyR3-licA* operon was consistent with the activity of this regulator as previously reported (11, 19), and each showed similar trends across all three growth stages (Fig. S3 and Table S3 in the supplemental material). The *orf4* gene (Clo1313_0398) showed 1.5 $\log_2$ fold higher expression in the *glyR3*

398 mutant only for stationary phase samples. Otherwise, no up regulation of the downstream genes
399 (*orf4-manB-celT*) was detected in the *glyR3* deletion strain similar to what was observed when
400 *C. thermocellum* was grown on laminarin (11). Clo1313_2013, annotated as encoding a
401 hypothetical protein, was the only other gene that maintained a greater than 2-fold up regulation
402 of gene expression across the three growth stages when *glyR3* was deleted. The most down-
403 regulated gene (Clo1313_0638) at mid-exponential and late-exponential growth phases
404 contained the same SNP as strain Δ*glyR2*. This was the only SNP in the coding region of a gene
405 in the Δ*glyR3* strain that displayed differential gene expression. Another gene with CAZyme
406 related functions is Clo1313_1424 (*celX*) encoding a predicted acetylxylan esterase cellulosomal
407 protein, which was up regulated by 2-fold in stationary phase. Two members of the free cellulase
408 system in *C. thermocellum* were up regulated by stationary phase, Clo1313_1336 (GH23) and
409 Clo1313_2473 (GH18) with possible substrates of chitin and/or peptidoglycan.
410
411 A histidine kinase (Clo1313_1711), predicted to be involved in regulating sporulation in
412 *C. thermocellum* was down regulated in late exponential and early stationary phase in the *glyR3*
413 deletion strain compared to the parental strain. Other genes with sporulation annotations
414 (Clo1313_0205, SpoIIID; Clo1313_1773, SigE; Clo1313_1774, SpoIIGA; Clo1313_1409
415 Spo0A; Clo1313_1548, Stage II Sporulation protein M; Clo1313_1381, Stage III Sporulation
416 protein AC; Clo1313_1381, Stage III Sporulation protein AF were down regulated at stationary
417 phase. The average read counts for the triplicate samples for many of these genes were low in
418 both the parental and mutant strain across all time points sampled. SpoIIID 3-49, SigE 123-1143,
419 SpoIIGA 12-55, Spo0A 128-648, Stage II sporulation protein M 5-35, Stage III Sporulation
420 protein AC 7-29, Stage III Sporulation protein AF 23-128. Spores were not identified during

culturing of the *lacI* mutants however *C. thermocellum* is a poor spore former, with rates of less than 10% under typical conditions that would induce sporulation and at less than 0.002% under typical lab conditions (40).

LacI regulators bind to regulatory sequences of target genes.

To test whether each LacI regulator has a direct role in the regulation of the genes that displayed increased expression in the transcriptomic analysis, each regulator was assessed for the ability to bind the DNA fragments from promoter regions of the genes of interest via several independent electrophoretic mobility shift assays (EMSA). Fluorescent EMSA was more sensitive and these results are presented. Chemiluminescent EMSA data were also collected using independent probe sequences (not shown). Purified LacI transcription factors were confirmed to be able to bind to DNA fragments from particular promoter regions, with increasing amounts of protein in the assay binding increasing amounts of DNA fragments (Fig. 4). GlyR3 bound to the region as previously described, GlyR1 bound upstream of *licB* (Clo1313_2022) and the GlyR2 regulator bound to a region upstream of *man5A* (Clo1313_1398).

LacI proteins typically dissociate from DNA upon binding of the effector molecule. While laminaribiose has been shown to be an effector molecule for GlyR3, effectors for GlyR1 and GlyR2 are unknown. Various sugars were tested with biotin labeled fragments to identify sugars that would disrupt binding. Cellobiose, laminaribiose (Fig 5), maltose, sucrose, glucose, lactose, galactose, arabinose, and xylose (data not shown) did not affect DNA binding by either GlyR1 or GlyR2, and α-mannobiose had no impact on GlyR1 binding. GlyR2, was no longer able to bind in the presence of α-mannobiose. This suggests that α-mannobiose is a probable inducer for

444 GlyR2 regulated genes (Fig. 5), consistent with the function of Clo1313_1398 as a mannanase.
445 Individual sugars did not yield a candidate inducer molecule for GlyR1, and the possible inducer
446 molecule for this transcription factor remains unknown.
447
448 **Discussion**
449 This study sought to identify the regulons of three LacI family transcription factors to better
450 understand regulation of genes involved in plant biomass deconstruction in *C. thermocellum*
451 DSM1313. The *C. thermocellum* genome encodes a large repertoire of enzymes, both cellulases
452 and hemicellulases that can target plant cell walls, degrading the cellulose and hemicellulose
453 components to liberate soluble β-1,4 and β-1,3 linked glucans that the bacterium will ferment.
454 The composition of the cellulosome has been shown to vary depending on the substrate (8) and
455 cellulase gene expression can be affected by the growth rate of the organism (5). However, the
456 regulatory mechanisms employed by the cell to determine the cellulosome composition and
457 *C. thermocellum* regulation in general remains relatively under studied.
458
459 LacI family members can sense sugar and non-carbohydrate effectors, with the majority
460 regulating carbohydrate utilization genes (23). A comparative study of 1,300 LacI transcription
461 factors found nearly 90% of these were local regulators of sugar utilization pathways, with
462 regulatory effects only impacting genes located near the regulator gene in the genome, while the
463 other 10% of LacI regulators control large sets of genes and are considered to be global
464 regulators. LacI family members can also be integrated into complex regulatory networks as is
465 the case for *Mesorhizobium loti* FixV, which activates genes for symbiotic nitrogen fixation (41)
466 as an example.

All three LacI transcription factors affected the regulation of genes encoding hemicellulases. Hemicelluloses constitute significant proportions of plant biomass polysaccharides, between a quarter to a half the mass of many types of cell walls, and they vary tremendously depending upon the plant species and tissue type (42). The four main kinds of hemicelluloses are xyloglucan, xylans, mixed-linkage glucans, and mannans, which contain different backbones with various side-groups, bonds, degrees of branching and polymerization. Xyloglucan is often the most abundant hemicellulose in primary cell walls and it consists of a β-1,4-glucan backbone with xylose-containing side chains, which may also contain other sugars (e.g. arabinose). However, in grass primary cell walls glucuronoarabinoxylan is a dominant hemicellulose and it consists of mixed linkage glucans. Xylan is a polymer of β-1,4-linked xylose with possible arabinan and glucuronic acid side chains. Mixed linkage glucans are made of glucose residues linked by β-1,3 and β-1,4 bonds. A mannose backbone joined by β-1,4 linkages make up mannans or similarly linked glucose and mannose residues comprise glucomannans. Wild-type *C. thermocellum* strains are unable to use these hemicellulose sugars as carbon sources. Nevertheless, *C. thermocellum* encodes a variety of enzymes needed to break hemicellulose bonds in order to access cellulose from a variety of sources and consume it for growth. We have identified new LacI repressor activity for genes with hemicellulase functions, which is a key result of this work.

The characterized ATCC 27405 genes *celC-glyR3-licA* (Cthe_2807, 2808, 2809) are homologous to DSM1313 genes Clo1313_0395, 0396, 0397. Expression studies across growth stages have identified a peak in the gene expression from this operon at late exponential as the cell transitions into early stationary phase (11, 43, 44). The activity of the GlyR3 regulator on this three gene

operon was demonstrated to have repressor activity, which was relieved in the presence of laminaribiose, a β-1,3-linked glucose dimer. GlyR3 was shown to bind to an 18 bp near-palindromic sequence upstream of the start codon of Cthe_2807 (19). GlyR3 has also been used to create a laminaribiose-inducible expression system (45). We find that all three LacI transcription factors in *C. thermocellum* DSM1313 appear to be local rather than global regulators and each LacI controlled the expression of a small number of genes encoding CAZyme functions. Deletion of each LacI transcription factor and transcriptome profiling across three different growth stages revealed the deletion of each transcription factor had a dramatic effect on the expression of at least one other gene across all time points sampled. This indicated that the absence of the regulator relieved the transcription repression and this was not confounded by the growth stage of the culture, consistent with roles as transcriptional repressors.

The regulon of GlyR3 appears to be limited to the *celC-glyR3-licA* operon, as previously described (11, 19). GlyR1 affected the expression of *licB* (Clo1313_2022), containing a glycoside hydrolase 16 domain. This gene encodes a lichenase (β-1,3-1,4-glucanase), which has been demonstrated to localize to the cellulosome (46) and cleaves β-1,4 linkages adjacent to β-1,3 linkages in mixed β-glucans, such as barley β-glucans and lichenen (47, 48). LicB has no activity against β-1,3- β-1,6 linkages such as laminarin (47) so has a function distinct from the LicA/CelC enzymes regulated by GlyR3. Purified GlyR1 bound to a 232 bp region upstream of the *licB* gene and this binding was dose dependent. Monomeric and multimeric sugar molecules were used to determine an inducer molecule for this regulator, however the binding of GlyR1 to the *licB* promoter region could not be disrupted and the inducer molecule remains unknown. Future studies that could fractionate different components of the lignocellulosic substrate or a

higher throughput assay would be useful to narrow down possible candidate effector molecules for GlyR1.

Deletion of *glyR2* led to significant up regulation of 17 genes (above 2-fold). Many had CAZyme activities and the expression of nine of these gene were previously reported to display a decreasing level of expression when the growth rate of *C. thermocellum* ATCC 27405 on cellobiose was increased (5). In the *glyR2* mutant strain, these genes had higher levels of expression at exponential growth phase compared to the parental strain and this trend was not observed in either of the other two strains, suggesting this effect is specific to this *lacI* deletion. Clo1313_1398, encoding Man5A, had substantially higher expression when *glyR2* was deleted. Interestingly, the regulation of Clo1313_1398 (*man5A*) in the two similar strains of *C. thermocellum* ATCC 27405 and DSM1313 appears to differ. Cthe_0821, the homolog of Clo1313_1398 is one of the most highly expressed genes when the cells were grown on biomass (7) and is in the top 6% of expressed genes grown on cellulose (5), while Clo1313_1398 is only in the top 30% of expressed genes when DSM1313 is grown on cellulose (2). The genomic architecture of this region does differ between the two strains, with DSM1313 lacking two transposase genes immediately upstream of Cthe_0821. This difference between the two strains has disrupted a σ^A recognition motif and might be the reason for the strain differences. EMSA assays revealed the purified GlyR2 bound to a 149 bp region upstream of the start site of the Clo1313_1398 gene in a dose dependent manner. This binding could be disrupted by the presence of mannobiose. This, along with the RNAseq data revealing high *man5A* expression levels in the *glyR2* deletion strain is indicative of the transcription factor acting as an inducible repressor. The ATCC27405 homolog displays endo-mannanase activity and over an extended

536 incubation period can hydrolyze mannopentaose to mannobiose (43). As mannobiose is both an
537 inducer and a product of the activity of Man5A, this suggests feedback regulation could occur in
538 the cell. Due to the genomic architecture of the region, this may be more important in the
539 DSM1313 strain than the ATCC 27405 strain where the gene is already expressed at a
540 constitutive level.
541
542 Deletions strains were resequenced, which identified a series of SNPs and other mutations
543 including transposon insertions in two of the strains. The occurrence of mutations and insertion
544 in phosphoribosyltransferases may result from selective pressure on nucleotide pathways during
545 the mutagenesis strategy employed to make these deletion strains. Fortunately, mutations were
546 identified and the effects associated with these mutations could be excluded from the
547 interpretation of the transcriptome analyses. It is recommended that mutants are routinely
548 resequenced to be aware of the genomic context of targeted deletions.
549
550 This study adds to the small body of knowledge for *C. thermocellum* gene regulation and
551 identified negative regulation affecting the expression of cellulosome related genes in this
552 bacterium. Previous studies have demonstrated negative regulation by GlyR3 transcription factor
553 of non-cellulosomal enzymes (11, 19). Positive regulation of cellulosome enzymes through
554 cellulose and xylan interactions with anti-σ factors has been reported (17, 20, 21). Further studies
555 are required to refine *C. thermocellum* gene-regulatory networks contributing to the control and
556 composition of the cellulosome and the lignocellulolytic abilities of the organism, regulation of
557 important physiological responses and cross-talk between networks.
558

Funding information

This work is supported by the BioEnergy Science Center (BESC), which is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. This manuscript has been authored by UT-Battelle, LLC under Contract no. DE-AC05-00OR22725 with the US Department of Energy. The funders had no role in study design, data collection and interpretation, preparation of the manuscript, or the decision to submit the work for publication.

Acknowledgments

The authors thank Evert Holwerda (Dartmouth) for fermentations suggestions. Miguel Rodriguez, Jr (ORNL) provided assistance with fermentations and HPLC systems.

References

1. **Blumer-Schuette SE, Brown SD, Sander KB, Bayer EA, Kataeva I, Zurawski JV, Conway JM, Adams MW, Kelly RM.** 2014. Thermophilic lignocellulose deconstruction. FEMS Microbiol Rev **38:**393-448.

2. **Xu Q, Resch MG, Podkaminer K, Yang S, Baker JO, Donohoe BS, Wilson C, Klingeman DM, Olson DG, Decker SR, Giannone RJ, Hettich RL, Brown SD, Lynd LR, Bayer EA, Himmel ME, Bomble YJ.** 2016. Dramatic performance of *Clostridium thermocellum* explained by its wide range of cellulase modalities. Sci Adv **2:**e1501254.
3. **Lynd LR, van Zyl WH, McBride JE, Laser M.** 2005. Consolidated bioprocessing of cellulosic biomass: an update. Curr Opin Biotechnol **16:**577-583.
4. **Nataf Y, Yaron S, Stahl F, Lamed R, Bayer EA, Scheper TH, Sonenshein AL, Shoham Y.** 2009. Cellodextrin and laminaribiose ABC transporters in *Clostridium thermocellum*. J Bacteriol **191:**203-209.
5. **Riederer A, Takasuka TE, Makino S-I, Stevenson DM, Bukhman YV, Elsen NL, Fox BG.** 2011. Global gene expression patterns in *Clostridium thermocellum* as determined by microarray analysis of chemostat cultures on cellulose or cellobiose. Appl Environ Microbiol **77:**1243-1253.
6. **Yaniv O, Fichman G, Borovok I, Shoham Y, Bayer EA, Lamed R, Shimon LJ, Frolow F.** 2014. Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of *Clostridium thermocellum* anti-sigmaI factors. Acta Crystallogr D Biol Crystallogr **70:**522-534.
7. **Wilson CM, Rodriguez M, Jr., Johnson CM, Martin SL, Chu TM, Wolfinger RD, Hauser LJ, Land ML, Klingeman DM, Syed MH, Ragauskas AJ, Tschaplinski TJ, Mielenz JR, Brown SD.** 2013. Global transcriptome analysis of *Clostridium thermocellum* ATCC 27405 during growth on dilute acid pretreated *Populus* and switchgrass. Biotechnol Biofuels **6:**179.
8. **Raman B, Pan C, Hurst GB, Rodriguez M, Jr., McKeown CK, Lankford PK, Samatova NF, Mielenz JR.** 2009. Impact of pretreated switchgrass and biomass carbohydrates on *Clostridium thermocellum* ATCC 27405 cellulosome composition: A quantitative proteomic analysis. PLoS One **4:**e5271.
9. **Bagnara-Tardif C, Gaudin C, Belaich A, Hoest P, Citard T, Belaich JP.** 1992. Sequence analysis of a gene cluster encoding cellulases from *Clostridium cellulolyticum*. Gene **119:**17-28.
10. **Kakiuchi M, Isui A, Suzuki K, Fujino T, Fujino E, Kimura T, Karita S, Sakka K, Ohmiya K.** 1998. Cloning and DNA sequencing of the genes encoding *Clostridium josui* scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. J Bacteriol **180:**4303-4308.
11. **Newcomb M, Millen J, Chen CY, Wu JH.** 2011. Co-transcription of the *celC* gene cluster in *Clostridium thermocellum*. Appl Microbiol Biotechnol **90:**625-634.
12. **Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, Wolf YI, Tatusov RL, Sabathe F, Doucette-Stamm L, Soucaille P, Daly MJ, Bennett GN, Koonin EV, Smith DR.** 2001. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol **183:**4823-4838.
13. **Sabathe F, Belaich A, Soucaille P.** 2002. Characterization of the cellulolytic complex (cellulosome) of *Clostridium acetobutylicum*. FEMS Microbiol Lett **217:**15-22.
14. **Schwarz WH.** 2001. The cellulosome and cellulose degradation by anaerobic bacteria. Appl Microbiol Biotechnol **56:**634-649.
15. **Tamaru Y, Doi RH.** 2000. The engL gene cluster of *Clostridium cellulovorans* contains a gene for cellulosomal *manA*. J Bacteriol **182:**244-247.

16. **Tamaru Y, Karita S, Ibrahim A, Chan H, Doi RH.** 2000. A large gene cluster for the *Clostridium cellulovorans* cellulosome. J Bacteriol **182:**5906-5910.
17. **Nataf Y, Bahari L, Kahel-Raifer H, Borovok I, Lamed R, Bayer EA, Sonenshein AL, Shoham Y.** 2010. *Clostridium thermocellum* cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. Proc Natl Acad Sci U S A **107:**18646-18651.
18. **Sander K, Wilson CM, Rodriguez M, Klingeman DM, Rydzak T, Davison BH, Brown SD.** 2015. *Clostridium thermocellum* DSM 1313 transcriptional responses to redox perturbation. Biotechnol Biofuels **8:**1-14.
19. **Newcomb M, Chen CY, Wu JH.** 2007. Induction of the *celC* operon of *Clostridium thermocellum* by laminaribiose. Proc Natl Acad Sci U S A **104:**3747-3752.
20. **Bahari L, Gilad Y, Borovok I, Kahel-Raifer H, Dassa B, Nataf Y, Shoham Y, Lamed R, Bayer EA.** 2011. Glycoside hydrolases as components of putative carbohydrate biosensor proteins in *Clostridium thermocellum*. J Ind Microbiol Biotechnol **38:**825-832.
21. **Kahel-Raifer H, Jindou S, Bahari L, Nataf Y, Shoham Y, Bayer EA, Borovok I, Lamed R.** 2010. The unique set of putative membrane-associated anti-sigma factors in *Clostridium thermocellum* suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. FEMS Microbiol Lett **308:**84-93.
22. **Mearls EB, Lynd LR.** 2014. The identification of four histidine kinases that influence sporulation in *Clostridium thermocellum*. Anaerobe **28:**109-119.
23. **Ravcheev DA, Khoroshkin MS, Laikova ON, Tsoy OV, Sernova NV, Petrova SA, Rakhmaninova AB, Novichkov PS, Gelfand MS, Rodionov DA.** 2014. Comparative genomics and evolution of regulons of the LacI-family transcription factors. Front Microbiol **5:**294.
24. **Papanek B, Biswas R, Rydzak T, Guss AM.** 2015. Elimination of metabolic pathways to all traditional fermentation products increases ethanol yields in *Clostridium thermocellum*. Metab Eng **32:**49-54.
25. **Argyros DA, Tripathi SA, Barrett TF, Rogers SR, Feinberg LF, Olson DG, Foden JM, Miller BB, Lynd LR, Hogsett DA, Caiazza NC.** 2011. High ethanol titers from cellulose by using metabolically engineered thermophilic, anaerobic microbes. Appl Environ Microbiol **77:**8288-8294.
26. **Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, McCool JD, Warner AK, Rajgarhia VB, Lynd LR, Hogsett DA, Caiazza NC.** 2010. Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a *pta* mutant. Appl Environ Microbiol **76:**6591-6599.
27. **Biswas R, Prabhu S, Lynd LR, Guss AM.** 2014. Increase in ethanol yield via elimination of lactate production in an ethanol-tolerant mutant of *Clostridium thermocellum*. PLoS ONE **9:**e86389.
28. **Biswas R, Zheng T, Olson DG, Lynd LR, Guss AM.** 2015. Elimination of hydrogenase active site assembly blocks H_2 production and increases ethanol yield in *Clostridium thermocellum*. Biotechnol Biofuels **8:**1-8.
29. **Tian L, Papanek B, Olson DG, Rydzak T, Holwerda EK, Zheng T, Zhou J, Maloney M, Jiang N, Giannone RJ, Hettich RL, Guss AM, Lynd LR.** 2016. Simultaneous achievement of high ethanol yield and titer in *Clostridium thermocellum*. Biotechnol Biofuels **9:**116.

30. **Lin PP, Mi L, Morioka AH, Yoshino KM, Konishi S, Xu SC, Papanek BA, Riley LA, Guss AM, Liao JC.** 2015. Consolidated bioprocessing of cellulose to isobutanol using *Clostridium thermocellum*. Metab Eng **31:**44-52.
31. **Maniatis T, Fritsch EF, Sambrook J.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
32. **Shanks RMQ, Kadouri DE, MacEachran DP, O'Toole GA.** 2009. New yeast recombineering tools for bacteria. Plasmid **62:**88-97.
33. **Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO.** 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods **6:**343-U341.
34. **Guss AM, Olson DG, Caiazza NC, Lynd LR.** 2012. Dcm methylation is detrimental to plasmid transformation in *Clostridium thermocellum*. Biotechnol Biofuels **5:**30.
35. **Olson DG, Lynd LR.** 2012. Transformation of *Clostridium thermocellum* by electroporation. Methods Enzymol **510:**317-330.
36. **Kridelbaugh DM, Nelson J, Engle NL, Tschaplinski TJ, Graham DE.** 2012. Nitrogen and sulfur requirements for *Clostridium thermocellum* and *Caldicellulosiruptor bescii* on cellulosic substrates in minimal nutrient media. Bioresour Technol **130:**125-135.
37. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods **9:**357-359.
38. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.
39. **Love MI, Huber W, Anders S.** 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15:**1-21.
40. **Mearls EB, Izquierdo JA, Lynd LR.** 2012. Formation and characterization of non-growth states in *Clostridium thermocellum*: spores and L-forms. BMC Microbiol **12:**180-180.
41. **Sullivan JT, Brown SD, Ronson CW.** 2013. The NifA-RpoN regulon of *Mesorhizobium loti* strain R7A and its symbiotic activation by a novel LacI/GalR-family regulator. PLoS ONE **8:**e53762.
42. **Mohnen D, Bar-Peled M, Somerville C.** 2009. Cell Wall Polysaccharide Synthesis, p 94-187. *In* Himmel ME (ed), Biomass Recalcitrance doi:10.1002/9781444305418.ch5. Blackwell Publishing Ltd.
43. **Mishra S, Beguin P, Aubert JP.** 1991. Transcription of *Clostridium thermocellum* endoglucanase genes *celF* and *celD*. J Bacteriol **173:**80-85.
44. **Raman B, McKeown CK, Rodriguez M, Jr., Brown SD, Mielenz JR.** 2011. Transcriptomic analysis of *Clostridium thermocellum* ATCC 27405 cellulose fermentation. BMC Microbiol **11:**134.
45. **Mearls EB, Olson DG, Herring CD, Lynd LR.** 2015. Development of a regulatable plasmid-based gene expression system for *Clostridium thermocellum*. Appl Microbiol and Biotechnol **99:**7589-7599.
46. **Zverlov VV, Fuchs KP, Schwarz WH, Velikodvorskaya GA.** 1994. Purification and cellulosomal localization of *Clostridium thermocellum* mixed linkage β-glucanase LicB (1,3-1,4-β-D-glucanase). Biotechnol Lett **16:**29-34.

47. **Schimming S, Schwarz WH, Staudenbauer WL.** 1991. Properties of a thermoactive beta-1,3-1,4-glucanase (lichenase) from *Clostridium thermocellum* expressed in *Escherichia coli*. Biochem Biophys Res Commun **177:**447-452.
48. **Schimming S, Schwarz WH, Staudenbauer WL.** 1992. Structure of the *Clostridium thermocellum* gene licB and the encoded beta-1,3-1,4-glucanase. A catalytic region homologous to *Bacillus* lichenases joined to the reiterated domain of clostridial cellulases. Eur J Biochem **204:**13-19.

Figure legends

Fig. 1. Growth of the parental and mutant strains on 5g.L^{-1} cellobiose in 1 L bioreactors for RNA-seq analysis. Plotted are average OD_{600} values from triplicate fermentations with error bars (within symbols and shown as standard deviations). Transcriptomics samples were taken at mid-log phase (OD_{600} of 0.7-0.95), late log phase (OD_{600} of 1.8-2.0) and in stationary phase approximately 30 min after cultures stopped producing acid (OD_{600} 2.0-2.2).

Fig. 2. RNA-Seq analysis for *man5A* region in Δ*glyR2* mutant and parent strains. ~30M reads for deletion (A) and parent (B) strains were mapped to the reference genome with the read numbers mapping indicated by coverage.

Fig. 3. Electrophoretic Mobility Shift Assays (EMSA) using purified LacI transcriptional regulators with DNA fragments from the promoters of regulated genes. Approximately 0.03µM, 0.02µM and 0.03µM of DNA fragments from the upstream genomic regions of Clo1313_2022, Clo1313_1398, Clo1313_0395 were incubated with increasing amounts of purified protein encoded by, Clo1313_2023 and Clo1313_0089 and Clo1313_0396 respectively. Lanes were loaded as: 100 bp marker (Life Technologies), DNA only, Lanes 3-11 DNA plus LacI regulator (GlyR1, GlyR2: 0.016, 0.03, 0.06, 0.13, 0.25, 0.50, 0.75, 1.0, 1.5 µM respectively), (GlyR3:

744 0.06, 0.13, 0.25, 0.50, 0.75, 1.0, 1.5, 2.0, 2,5 μM) Lane 12 Protein only. FP indicates where the
745 free probe is running on the gel.

746

747 Fig. 4. Electrophoretic mobility shift assays (EMSA) using purified LacI transcriptional
748 regulators with biotin labeled DNA fragments from the promoters of regulated genes. 20 fmol of
749 biotin labeled probe and 15 nM of purified transcription factor. Control competitive binding
750 reactions included 4 pmol of unlabelled probe. Bound probe (BP), free probe (FP) are indicated
751 and sugars were included in experiments at designated quantities.

752

753 **Table 1. Oligonucleotides used in this study.**

Primer Name	Sequence
MUTANT CONSTRUCTION	
Cthe0210up-f-4AMG550Bam	TGATCCACTAGTAAGCTTACCCCTGCGAATACTGAGATTT
Cthe0210up-r-for0210down	GATTTTAAGTTGAGACAAATCAGTTTTCTTTGTATTT
Cthe0210down-f-for0210up	ATTTGTCTCAACTTAAAATCAGCCTTTTAAGTAAAGAA
Cthe0210down-r-4AMG550Bam	GGCCTCGCGAAGGAGTTCTTCCGTCGTAACGT
Cthe0210in-f-4AMG550Eco	CCTTTTGTTTTGGTACCGTCAAGCAATAATTCAATCGAGC
Cthe0210in-r-4AMG550Eco	AAACGCTGAGGCGCGCCGATCGTCGGAGTTTCAAGAAG
Cthe2505up-f-4AMG550Bam	GATCCACTAGTAAGCTTGCCTATGGTTCCGCAACTTTC
Cthe2505up-r-for2505down	TCAGCCTATTCGCTTTACTGACATTGTGACTTTCTTTGCCATT
Cthe2505down-f-for2505up	AAGAAAGTCACAATGTCAGTAAAGCGAATAGGCTGAT
Cthe2505down-r-4MG550Bam	ACTGGCCGGCCTCGCGAGTTGGGAACATTTCTGTGAATTC
Cthe2505in-f-4AMG550EcoRI	CCTTTTGTTTTGGTACCGAGCAAGTCAAATTCAAGAACCG
Cthe2505in-r-4AMG550EcoRI	AAACGCTGAGGCGCGCCGTCATCGGGAACGGATATACC
Ct2808down-r-4AMG550	ATTGATACACCCATGTCAGAATACTGGCCGGCCTCGCGAGGATCCGATTTAA CGACAACACTTTCAACC
Ct2808down-f-42808up	TAATTTAAAAAATACAACAGCTGTTGATTGATTGTTTTATAATCATCCAAAGA TGTACTC
Ct2808up-f- 4AMG550	GGCGCTACAGGGCGCGTGGGGATGATCCACTAGTAAGCTTGGATCCATTAC GGAGAAGGACATTGAAACTATT
Ct2808up-r- 42808down	ACATCTTTGGATGATTATAAAACAATCAATCAACAGCTGTTGTATTTTTTAAA TTACAAA
Ct2808in-f- 4AMG550(SmaI)	TCCTTTTATAGGCGTTTTGAAACCTGAAATGTACAGCTAAGGAGGTGGGCCC ATGACCAGTGAAGAAATAGCAAAATTAT
Ct2808in-r- 4AMG550(SmaI)	AAGGCCCAGTCTTCCGACTGAGCCTTTTGTTTTCTCGAGGTCAGAATTCCAAA GCCCTCTTGGT
MUTANT VERIFICATION	

F_AAACCTGTCTCGCAAGAATTG	AAACCTGTCTCGCAAGAATTG (Upstream of Che_2808)
R_CGCGTTTCCTCTTTTACGTT	CGCGTTTCCTCTTTTACGTT (Reverse within Cthe_2808)
R_GTACGGATATTGGCGACCAG	GTACGGATATTGGCGACCAG (Downstream at end of Cthe_2809)
F_CTTGTTTACCGCATCTGCAA	CTTGTTTACCGCATCTGCAA (5' of Cthe_2505)
R_AATTCCTCAACTGCCCACAG	AATTCCTCAACTGCCCACAG (Reverse within Cthe_2505)
R_TTTGCCTGCATTACAACAGC	TTTGCCTGCATTACAACAGC (3' of Cthe_2505)
F_TTTCATGGCCTTCATGTTGA	TTTCATGGCCTTCATGTTGA (5' of Cthe_0210)
R_CATGGCTATCGGAGCATACA	CATGGCTATCGGAGCATACA (Reverse within Cthe_0210)
R_AACACTTCAGCAAAGCTCCTG	AACACTTCAGCAAAGCTCCTG (3' of Cthe_0210)
CLONING	
F_pTXB1 Clo1313_2023/0210	TAATTTTGTTTAACTTTAAGAAGGAGATATACAATGAACAGCAAGGATATAGC
R_pTXB1_CLO1313_2023/0210	GGGTAGGGCAACTAGTGCATCTCCCGTGATGCAGACAATTTTTTTACATGAATTACGC
F_pTXB1_CLO1313_0089/2505	TAATTTTGTTTAACTTTAAGAAGGAGATATACAATGGCAAAGAAAGTCACAATG
R_pTXB1_Clo1313_0089/2505	GGGTAGGGCAACTAGTGCATCTCCCGTGATGCATATTCGCTTTACTGAC
F_pTXB1_Clo1313_0396/2808	TAATTTTGTTTAACTTTAAGAAGGAGATATACAATGACCAGTGAAGAAATAGC
R_pTXB1_Clo1313_0396/2808	GGGTAGGGCAACTAGTGCATCTCCCGTGATGCAGAATTCCAAAGCCCTCTTG
F_pTXB1 across insertion	TAATACGACTCACTATAGGG
R_pTXB1 across insertion	GTAGGGCAACTAGTGCATCT
R. pTXB1 amp up of insertion	TGTATATCTCCTTCTTAAAG
F. pTXB1 amp down of insertion	ATCACGGGAGATGCACTAGT
FLUORESCENT EMSA	
F_EMSA_288bp_2022	GCAAAGCTCCTGTAACAATTCA
R_EMSA_288bp_2022	TGCCACACCACCTTCATATCA
F_EMSA_100bp_0395	CCGAATAAAAACTGGACAGAG
R_EMSA_100bp_0395	TGAAACCATTTAACACTGGATTAT
F_EMSA_refined_int1397:98	TTCAATAAAGCGGAATTTGAAGA
R_EMSA in refined 1398	CGGAGGACTTGTGTCGGTT
CHEMILUMINESCENT EMSA	
F_EMSA refined_int1397:98Biotin	TTCAATAAAGCGGAATTTGAAGA
R_EMSA in refined 1398	CGGAGGACTTGTGTCGGTT
F_EMSA 288bp_2022	GCAAAGCTCCTGTAACAATTCA
R_EMSA_2022_6_Biotin	TAGTTGACAATCTGAAGTTTA
SNP DETECTION	
F_Clo1313_0908_1058322_1058	CAAGGAGGATTTAACATGGATTT
R_Clo1313_0908_1058322_1058	ACCACCCTTTGACCCTTTTT
F_Clo1313_1891_2208653_2208	TAGCATCCGGACTGGTTTTT
R_Clo1313_1891_2208653_2208	ATGAATGAAAACAATCTAAATA
F_Clo1313_0145_159861_15986	CAGATCCCAGTTGGTTTGGA

R_Clo1313_0145_159861_15986	CAGGAAAATACAGGCTGTTGG
F_Clo1313_0185_199883_19988	GTTGGCAAAATCGGTGAGAT
R_Clo1313_0185_199883_19988	CATCGGATCGAGCACTACAA

754
755
756

757 **Table 2. Summary of significant gene expression differences (+/- 2-fold and adj. p<0.05).**

Growth stage of sampling	Summary of gene expression differences	GlyR1: ΔClo1313_2023 vs. Parental strain Δ*hpt*	GlyR2: ΔClo1313_0089 vs. Parental strain Δ*hpt*	GlyR3: ΔClo1313_0396 vs. Parental strain Δ*hpt*
Mid-exponential	Up regulated by 2-fold	2	17	8
	Down regulated by 2-fold	11	24	23
Late exponential	Up regulated by 2-fold	20	8	23
	Down regulated by 2-fold	13	18	42
Stationary	Up regulated by 2-fold	572	8	30
	Down regulated by 2-fold	520	38	45

758
759

Table 3. Summary of major gene expression differences in the three *lacI* deletion strains.

				Differential expression (log_2)		
Locus Tag	Function	SNP (Y/N)	Growth phase	*ΔglyR1* –Δ*hpt*	*ΔglyR2*– Δ*hpt*	*ΔglyR3*– Δ*hpt*
Clo1313_0089	GlyR2	N	Mid log	-0.01	Del.	0.09
			Late Log	-0.01	Del.	-0.19
			Early St	-1.47	Del.	0.09
Clo1313_0395	CelC	N	Mid log	-0.12	-0.29	3.89
			Late Log	0.03	-0.25	3.82
			Early St	-0.65	-0.21	2.53
Clo1313_0396	GlyR3	N	Mid log	0.12	-0.38	Del.
			Late Log	0.44	-0.09	Del.
			Early St	-0.67	-0.38	Del.
Clo1313_0397	LicA	N	Mid log	-0.09	-0.19	3.82
			Late Log	0.01	-0.08	3.75
			Early St	-0.18	-0.07	2.33
Clo1313_1398	Man5A	N	Mid log	-0.05	4.65	0.06
			Late Log	0.02	4.58	0.06
			Early St	-0.08	5.66	0.04
Clo1313_2022	LicB	N	Mid log	4.33	0.70	0.8
			Late Log	4.44	0.47	0.39
			Early St	6.64	0.22	0.29
Clo1313_2023	GlyR1	N	Mid log	Del.	-0.49	-0.27
			Late Log	Del.	-0.22	-0.29
			Early St	Del.	0.11	0.21

Del. indicates the gene was deleted in that particular strain and Stationary is abbreviated, St.

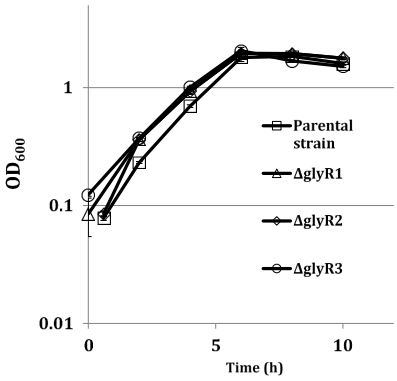
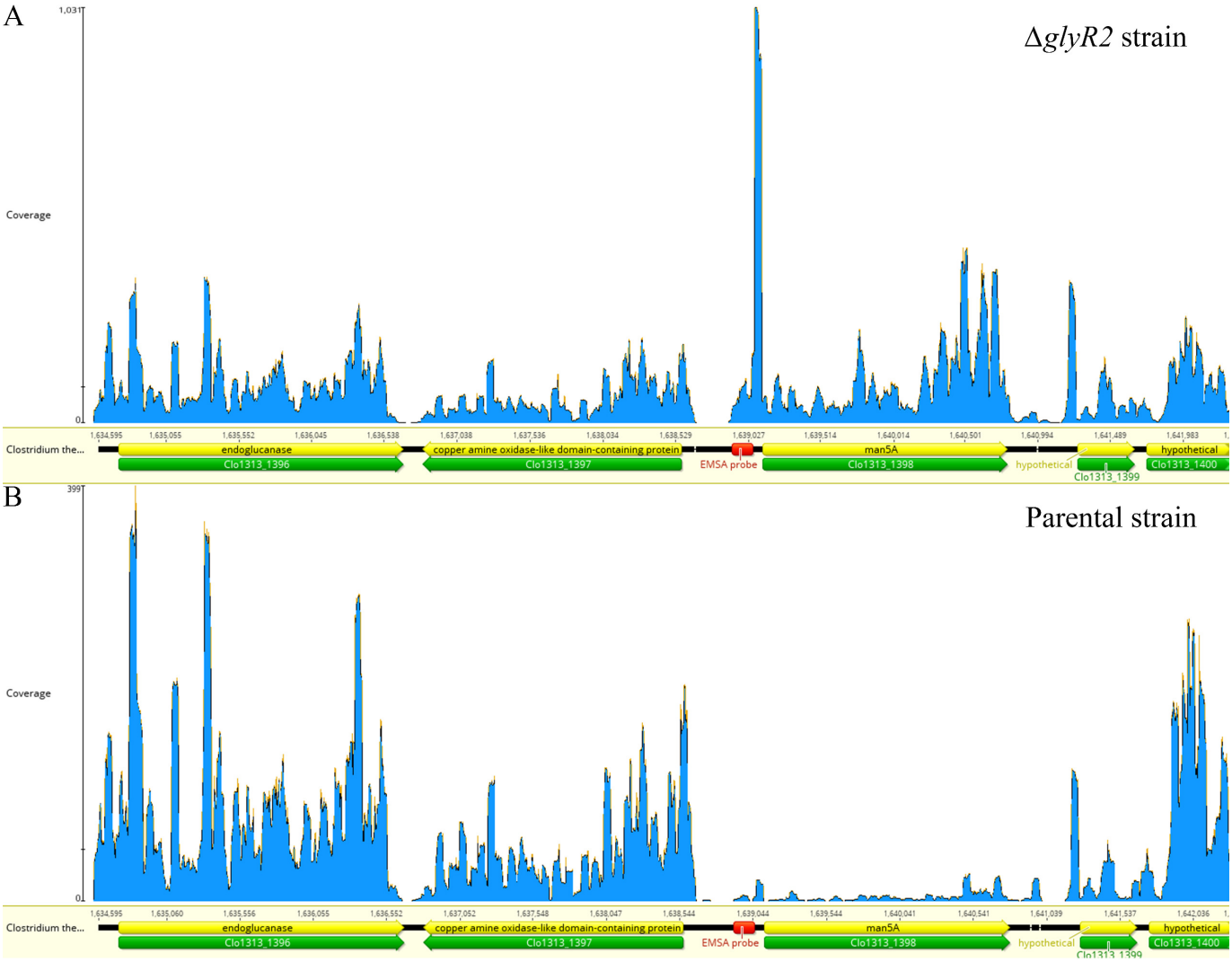


Figure 1



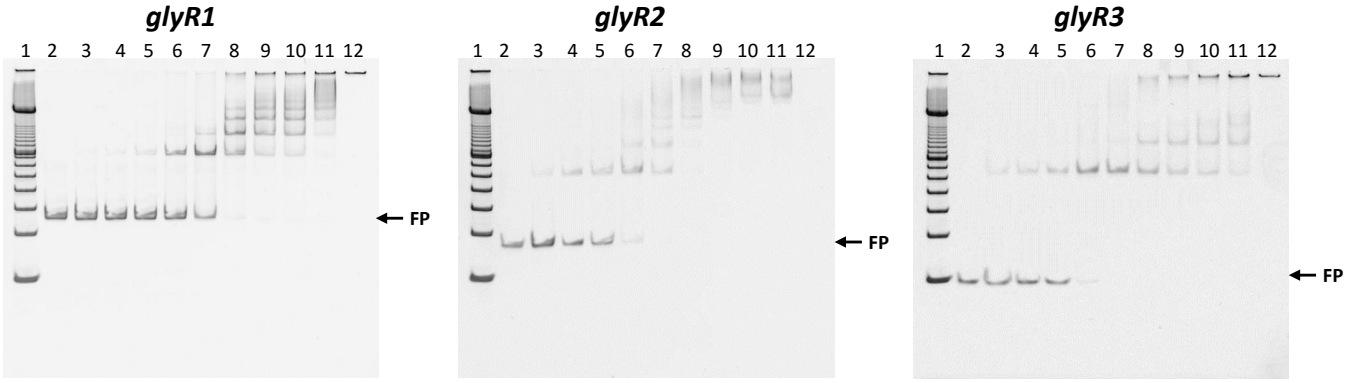


Figure 3

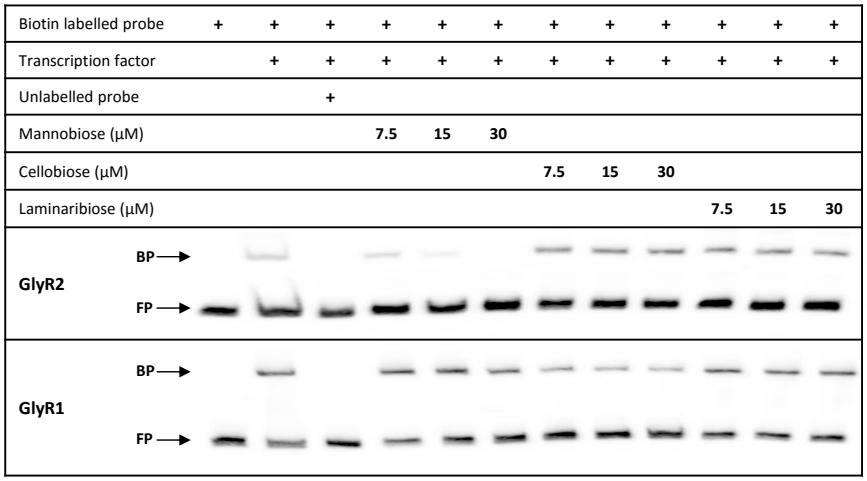


Figure 4