

1 Nov 13, 2018 rmk

2

3 **Comparative Biochemical and Structural Analysis of Novel Cellulose Binding**  
4 **Proteins (Tāpirins) from Extremely Thermophilic *Caldicellulosiruptor* Species**

---

5

6 Laura L. Lee<sup>1</sup>, William S. Hart<sup>1</sup>, Vladimir V. Lunin<sup>2</sup>, Markus Alahuhta<sup>2</sup>, Yannick J. Bomble<sup>2</sup>,  
7 Michael E. Himmel<sup>2</sup>, Sara E. Blumer-Schuetz<sup>1,^</sup>, Michael W.W. Adams<sup>3</sup>, and Robert M. Kelly<sup>1,\*</sup>

8

9 <sup>1</sup>*Department of Chemical and Biomolecular Engineering*  
10 *North Carolina State University, Raleigh, NC 27695*

11

12 <sup>2</sup>*Biosciences Center, National Renewable Energy Laboratory, Golden, CO, 80401*

13

14 <sup>3</sup>*Department of Biochemistry and Molecular Biology*  
15 *University of Georgia, Athens, Georgia 30602*

16

17 <sup>^</sup>*Present address: Department of Biological Sciences, Oakland University, Rochester, MI*

18

19 **Submitted to: *Appl Environ Microbiol* (August, 2018)**

20

**Revision to AEM01983 submitted November, 2018**

21

22 **Keywords:** *Caldicellulosiruptor*, tāpirins, lignocellulose, glycoside hydrolase, cellulase

23

24 **Running title:** Novel binding proteins in *Caldicellulosiruptor*

25

26

27

28 \*Address correspondence to:

29

30

31

32

33

34

35

36

**Robert M. Kelly**

Dept. of Chemical and Biomolecular Engineering

North Carolina State University

EB-1, 911 Partners Way

Raleigh, NC 27695-7905

Phone: (919) 515-6396

Fax: (919) 515-3465

Email: [rmkelly@ncsu.edu](mailto:rmkelly@ncsu.edu)

37 **Abstract**

38

39 Genomes of extremely thermophilic *Caldicellulosiruptor* species encode novel cellulose binding  
40 proteins, tāpirins, located proximate to the type IV pilus locus. The C-terminal domain of a  
41 tāpirin (Calkro\_0844) from *Caldicellulosiruptor kronotskyensis* is structurally unique and has a  
42 cellulose binding affinity akin to family 3 carbohydrate binding modules (CBM3). Here, full-length  
43 and C-terminal versions of tāpirins from *Caldicellulosiruptor bescii* (Athe\_1870),  
44 *Caldicellulosiruptor hydrothermalis* (Calhy\_0908), *Caldicellulosiruptor kristjanssonii*  
45 (Calkr\_0826), and *Caldicellulosiruptor naganoensis* (NA10\_0869) were produced recombinantly  
46 in *Escherichia coli* and compared to Calkro\_0844. All five tāpirins bound to microcrystalline  
47 cellulose, switchgrass, poplar, filter paper, but not to xylan. Densitometry analysis of bound  
48 protein fractions visualized by SDS-PAGE revealed that Calhy\_0908 and Calkr\_0826 (from  
49 weakly cellulolytic species) associated with the cellulose substrates to a greater extent than  
50 Athe\_1870, Calkro\_0844 and NA10\_0869 (from strongly cellulolytic species). Perhaps this  
51 relates to their specific needs to capture glucans released from lignocellulose by cellulases  
52 produced in *Caldicellulosiruptor* communities. Calkro\_0844, and NA10\_0869 share a high  
53 degree of amino acid sequence identity (> 80% identity), more so than with Athe\_1870 (~50%).  
54 The amino acid sequence identities of Calhy\_0908 and Calkr\_0826 compared to Calkro\_0844  
55 were only 16% and 36%, respectively, although the three-dimensional structures of their C-  
56 terminal binding regions were closely related. Unlike the parent strain, *C. bescii* mutants lacking  
57 the tāpirin genes did not bind to cellulose following short-term incubation, suggesting a role in  
58 cell association with plant biomass. Given the scarcity of carbohydrates in neutral terrestrial hot  
59 springs, tāpirins likely help scavenge carbohydrates from lignocellulose to support growth and  
60 survival of *Caldicellulosiruptor* species.

61

62 **Importance**

63  
64 Mechanisms by which microorganisms attach to and degrade lignocellulose are important to  
65 understand if effective approaches for conversion of plant biomass into fuels and chemicals are  
66 to be developed. *Caldicellulosiruptor* species grow on carbohydrates from lignocellulose at  
67 elevated temperatures and have biotechnological significance for that reason. Novel cellulose  
68 binding proteins, called tāpirins, are involved in the way *Caldicellulosiruptor* species interact with  
69 microcrystalline cellulose and here additional information about the diversity of these proteins  
70 across the genus is provided, including binding affinity and three-dimensional structural  
71 comparisons.

## Introduction

72  
73  
74  
75 The natural capacity to utilize both the cellulose and hemicellulose content of plant  
76 biomass as microbial growth substrates is relatively rare, especially among extreme  
77 thermophiles growing optimally above 70°C (1). However, in pH neutral, terrestrial hot springs  
78 and thermal features, species from the genus *Caldicellulosiruptor* can be isolated, all of which  
79 utilize hemicellulose, but only some of which can hydrolyze microcrystalline cellulose (2, 3). To  
80 degrade plant material, *Caldicellulosiruptor* species draw from an inventory of intracellular,  
81 surface (S)-layer associated, and secreted glycoside hydrolases (GHs) with complementary  
82 modes of action (4, 5). Cellulosomal or free enzyme systems, which can be modular, are  
83 commonly found in other cellulolytic organisms, such as *Clostridia* (6) and *Trichoderma* (7),  
84 respectively. Many *Caldicellulosiruptor* carbohydrate active enzymes (CAZymes (8)) are also  
85 modular, consisting of combinations of catalytic and non-catalytic (e.g., carbohydrate binding  
86 module [CBM]) domains connected by proline/threonine-rich linkers in various arrangements.  
87 The most well studied example is the multi-functional cellulase, CelA, which is arranged as  
88 GH9-CBM3-CBM3-CBM3-GH48 domains, where the numbers refer to specific protein families  
89 (9-15). The synergy between the endoglucanase (GH9) and exoglucanase (GH48) domains  
90 contributes to a novel mode of action for CelA that involves physical burrowing into cellulose  
91 fibers, thereby creating cavities for further enzymatic access to the carbohydrate content of  
92 plant biomass (9).

93 While not directly responsible for lignocellulose degradation, the non-catalytic domains in  
94 CelA also play an important role. In general, CBMs improve the efficacy of GHs by ensuring  
95 proximity to the substrate, as well as in some cases contributing to thermostability (16, 17).  
96 CBM3s, in particular, are specific to cellulose and allow enzymes like CelA to attach to their  
97 substrates such that their GH domains are proximate to their substrate (9). Other non-catalytic  
98 protein features in the *Caldicellulosiruptor* also play a role in orienting cells to lignocellulosic

99 carbohydrates. S-layer homology (SLH) domains are associated with certain modular GHs in  
100 these bacteria (18, 19). For instance, Calkro\_0402, a xylanase with GH10, CBM22, and CBM9  
101 domains, is anchored to the cell surface of the strongly cellulolytic *Caldicellulosiruptor*  
102 *kronotskyensis* and the gene encoding this enzyme is highly transcribed during growth on  
103 lignocellulose (switchgrass). When inserted into the genome of *Caldicellulosiruptor bescii*,  
104 Calkro\_0402 improved the attachment of cells to xylan and significantly increased xylan  
105 degradation, despite the fact that wild type *C. bescii* produces other xylanases (18).

106 In addition to CBMs within modular enzymes, *Caldicellulosiruptor* species also use non-  
107 catalytic proteins to bind to lignocellulosic substrates. Transcriptomic and proteomic analysis of  
108 cellulose-bound *Caldicellulosiruptor* cultures identified the presence of carbohydrate binding  
109 proteins (20), including the recently characterized 'tāpirins' (21). Tāpirins typically have a  $M_r$  of  
110 approximately 70 kDa as a single polypeptide; recombinant versions from *C. kronotskyensis*  
111 have a specific affinity for cellulose fibers in plant material and an affinity for Avicel similar to  
112 that of CBM3s, despite no significant structural homology (21). While the full-length structure  
113 has not been resolved, the cellulose-binding, 38.4 kDa C-terminal domain from Calkro\_0844  
114 was successfully crystallized and shown to be novel within the current protein database.  
115 Hydrophobic and aromatic residues present on the face of a  $\beta$ -helix likely make up the binding  
116 pocket with a flexible loop overhanging, and potentially protecting, access to it (21).  
117 Interestingly, tāpirin genes are located near the type IV pilus (T4P) locus in *Caldicellulosiruptor*  
118 species, suggesting a potential functional connection.

119 In the present study, comparative assessment of tāpirins across the genus  
120 *Caldicellulosiruptor* was conducted (see **Figure 1**). Structural data for two additional tāpirins  
121 from less cellulolytic species are provided, as is an assessment of relative binding capacities.  
122 Additionally, the role of the tāpirins was further explored through gene deletions in *C. bescii* and  
123 the resulting impact on binding.

124

125

## Results and Discussion

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

**Tāpirins are necessary for rapid binding to cellulosic substrates.** Degradation of

lignocellulosic substrates by *Caldicellulosiruptor* populations is likely predicated by substrate

attachment, as genes encoding for key cellulases from the Glucan Degradation Locus (GDL)

(22) are genomic neighbors of a T4P locus and tāpirins (**Figure 1**) (21). To assess the

importance of the tāpirins and/or T4P in binding to cellulose, knockouts of tāpirins (Athe\_1870-

1871) and the entire pilus locus plus tāpirins (Athe\_1870-1885) were generated in *C. bescii*.

After an hour incubation with Avicel, unlike the parent strain, both knockouts showed no

propensity for binding, based on changes on planktonic cell densities (**Figure 2**). Statistically

insignificant changes were noted in the planktonic cell density for the tāpirins and tāpirins/T4P

KOs when grown in the presence of microcrystalline cellulose, while in the parent strain

planktonic cell density was reduced by more than  $7 \times 10^7$  cells/ml. This suggests that the

tāpirins play a role in cell adherence to cellulosic substrates, at least during initial exposure,

which could be critical for scavenging carbohydrates in otherwise nutrient-limited hot springs.

**Tāpirins are ubiquitous in the genus *Caldicellulosiruptor*.** Putative tāpirins are found

in all genome-sequenced *Caldicellulosiruptor* species and belong to one of two groups (Class 1

and Class 2), based on their localization with genes encoding the Glucan Degradation Locus

and T4P; class 1 tāpirins are closest to the T4P locus and class 2 tāpirins (if there are more

than one tāpirin) closest to the GDL (**Figure 1**) (21). The GDL encodes up to seven glycoside

hydrolases (GHs), which collectively contain GH5, GH9, GH10, GH12, GH44, GH48, GH74, and

CBM3 domains, and these enzymes play an essential role in microcrystalline cellulose

hydrolysis (22). The most cellulolytic *Caldicellulosiruptor* species (e.g., *C. bescii*, *C.*

*kronotskyensis*, and *C. naganensis*) produce six to seven GDL GHs, while genomes from the

less cellulolytic species possess an incomplete set of the GDL enzymes, lacking modular

151 enzymes with a GH48 domain. For example, *C. kristjanssonii*, which is less cellulolytic,  
152 produces two of the GDL enzymes, while *C. hydrothermalis*, which minimally degrades  
153 microcrystalline cellulose, lacks all of the GDL GHs (**Figure 1**). Amino acid sequence analysis  
154 indicates that while some tāpirins are closely related (e.g., Calkro\_0844 and NA10\_0869 from  
155 two highly cellulolytic species, *C. kronotskyensis* and *C. naganoensis*, are 85% identical at the  
156 amino acid level), Calhy\_0908 from the weakly cellulolytic *C. hydrothermalis* is less than 18%  
157 identical to the other tāpirins. Interestingly, the *N*-termini of these tāpirins appear to share a high  
158 number of identical and conserved amino acids across the whole protein sequence (seen with  
159 residues 1-298 in **Figure S1**, on the right side of the indicated linker), with the exception of  
160 Calhy\_0908; in fact, Athe\_1870 shares 86% amino acid identity in this *N*-terminal range with  
161 both Calkro\_0844 and NA10\_0869 (**Table S1A**). Overall, the *N*-terminus of the tāpirin may be  
162 responsible for how tāpirins, in general, associate with the cell surface, while the *C*-terminus  
163 (%ID in **Table S1B**) establishes the binding function, the latter of which was determined to be  
164 the case for Calkro\_0844 (21).

165 ***In vitro* binding assays with tāpirins and plant-based substrates.** Tāpirins were  
166 initially characterized from *C. kronotskyensis* (Calkro\_0844 and Calkro\_0845) and from  
167 *Caldicellulosiruptor saccharolyticus* (Csac\_1073) (21). Binding assays showed that tāpirins from  
168 these two species preferentially adhered to cellulose. To confirm that this substrate specificity  
169 was consistent across the genus *Caldicellulosiruptor*, four additional tāpirins from weakly to  
170 strongly cellulolytic species were recombinantly produced to examine their binding to plant  
171 biomass-related substrates (see **Figure 3**). Athe\_1870, Calhy\_0809, Calkr\_0826, and  
172 NA10\_0869 from *C. bescii*, *C. hydrothermalis*, *C. kristjanssonii*, and *C. naganoensis*,  
173 respectively, along with the previously characterized Calkro\_0844, were incubated with Avicel,  
174 filter paper and xylan, as well as with lignocellulose (switchgrass and poplar). After incubation,  
175 the samples were split into 'unbound' and 'substrate-bound' fractions, where the 'substrate-

176 bound' protein was released upon denaturation in Laemmli sample buffer, after which both  
177 fractions were visualized with SDS-PAGE. No tāpirins bound to xylan to any significant extent,  
178 but consistently adhered to the other substrates. It was interesting that tāpirins from weakly  
179 cellulolytic species (Calkr\_0826 and Calhy\_0908) bound to poplar and switchgrass to a greater  
180 extent than the tāpirins from more strongly cellulolytic species. In fact, the tāpirin from the least  
181 cellulolytic species tested, Calhy\_0908, appears to adhere to more binding sites on cellulosic  
182 materials overall.

183 Densitometry analysis of the gels from the tāpirin binding assays supported the  
184 conclusions reached from visual inspection (see **Figure 4**). All tāpirins tested had a binding  
185 preference for purified cellulose (filter paper more so than Avicel, despite their similar  
186 crystallinity (23)). The larger particle size of filter paper (5/16" wide circular disks) compared to  
187 Avicel (50  $\mu\text{m}$  particles), in addition to higher protein absorption capacity (24, 25), may have  
188 been responsible for this difference. Interestingly, even though the tāpirins from highly  
189 cellulolytic *C. kronotskyensis* and *C. naganoensis* share 85% amino acid identity, NA10\_0869  
190 bound equally as well to switchgrass as it did to purified cellulose, in contrast to Calkro\_0844.  
191 Aside from *C. naganoensis*, overall tāpirin binding to lignocellulosic substrates was less than  
192 that for filter paper and Avicel, likely because of inaccessibility to microcrystalline cellulose in the  
193 plant biomasses. As indicated by the density of SDS-PAGE bands, both tāpirins from weakly  
194 cellulolytic species (Calhy\_0908 and Calkr\_0826) bound better to Avicel, poplar, switchgrass  
195 and filter paper. Calhy\_0908 bound better to all substrates compared to Calkr\_0826 except for  
196 switchgrass (**Figure 3**). In fact, the bands of Calhy\_0908 and Calkr\_0826 are approximately 3  
197 times darker than Athe\_1870 and range from 10- to 13-fold more intense than Calkro\_0844 and  
198 NA10\_0869 on filter paper (**Table S2**). Similar trends are noted, but to a lesser extent, on  
199 Avicel, with more Calhy\_0908 bound than Calkr\_0826 and Athe\_1870 (2- to 3-fold) and  
200 significantly more than NA10\_0869 (8-fold) and Calkro\_0844 (16-fold) (**Table S2**). It is possible  
201 that since the less cellulolytic *Caldicellulosiruptor* species, such as *C. kristjanssonii* and *C.*

202 *hydrothermalis*, cannot hydrolyze cellulose as well as other species (2), proximity to these  
203 substrates allows these species to exploit the collective hydrolytic capacity of cellulolytic  
204 communities in their natural environments.

205 **Structural comparisons of tāpirins from *Caldicellulosiruptor hydrothermalis*,**  
206 ***Caldicellulosiruptor kristjanssonii*, and *Caldicellulosiruptor kronotskyensis*.** The C-  
207 terminal domains of tāpirins from *C. hydrothermalis* and *C. kristjanssonii*, Calhy\_0908C and  
208 Calkr\_0826C, respectively, exhibited the same structural architecture as Calkro\_0844C from *C.*  
209 *kronotskyensis*, with the core of the domain being a  $\beta$ -helix with a characteristic long loop  
210 connecting the ends of the helix (**Figure 5C**). This fold was observed in the Calkro\_0844C  
211 structure (21) and seems to be a common fold for tāpirins.

212 The Calkr\_0826C structure was superimposed onto Calkro\_0844C with the root-mean-  
213 square deviation (r.m.s.d.) of 0.854 Å over 1,442 atoms, with both domains having the core  $\beta$ -  
214 helix of the same size (**Figure 5A**). The  $\beta$ -helix in both cases has 11 complete turns, with the  
215 longest  $\beta$ -sheet (designated face A, see **Figure 5D**) having 14  $\beta$ -strands. A few differences,  
216 however, are evident: a loop between  $\beta 8$  and  $\beta 9$  in Calkro\_0844C is 6 residues longer,  $\beta$ -strand  
217  $\beta 26$  is missing in Calkro\_0844C, and the  $\alpha$ -helix in Calkro\_0844C before  $\beta 27$  is not present in  
218 Calkr\_0826C. The long loop connecting the ends of the  $\beta$ -helix ( $\beta 28$  to  $\beta 29$ , Calkro\_0826C) is of  
219 the same length (40 residues) in both Calkr\_0826C and Calkro\_0844C, and features an  $\alpha$ -helix  
220 located at the same position in the middle. The loop connecting the  $\alpha 2$  helix to the  $\beta 32$  strand is  
221 3 amino acid residues shorter in Calkr\_0826C, which is compensated by the neighboring loop  
222 between  $\beta 33$  and  $\beta 34$ , which is 9 residues longer in Calkr\_0826C. Also, the loop between  $\beta 36$   
223 and  $\beta 37$  is again 3 residues shorter in Calkr\_0826. It is worth noting that the connecting loops of  
224 different lengths ( $\beta 8$ - $\beta 9$ ,  $\alpha 2$ - $\beta 32$ ,  $\beta 33$ - $\beta 34$  and  $\beta 36$ - $\beta 37$ ) are adjacent and represent about half  
225 of the edge between faces B (the next one after A following the direction of the polypeptide, see  
226 **Figure 5D**) and C (the face following B, see **Figure 5D**) of the  $\beta$ -helix, opposite to the

227 connecting loop. Similar to the Calkro\_0844C, Calkr\_0826C has a hydrophobic surface on the  
228 face A covered by the connecting loop with multiple flat-on-a-surface aromatic sidechains lined  
229 up along the  $\beta$ -sheet (**Figure 5A**).

230 The C-terminal domain of Calhy\_0908C is longer than that of Calkro\_0844C and shares  
231 the least amount of amino acid sequence homology (**Table S1**) to the other tãpirins examined  
232 here. The lack of homology actually led to an incorrect sequence alignment by NCBI-BLAST  
233 (26), which in turn translated into an incorrect homology model that was unsuccessfully  
234 attempted for molecular replacement. When the molecular replacement attempt failed, the  
235 structure of the Calhy\_0908C was determined via single-wavelength anomalous diffraction  
236 (SAD) using the anomalous signal of iodine atoms incorporated into the crystal after a short  
237 soak. Once the structure was determined, a structure-based sequence alignment was a more  
238 reliable basis for comparing different tãpirins (**Figure 5B-C**).

239 Calhy\_0908C exhibits the same overall architecture as the other two tãpirins with known  
240 structures (**Figure 5C-D**): the  $\beta$ -helix as the core of the domain and a long loop connecting the  
241 ends on the  $\beta$ -helix. The major difference between the Calhy\_0908C and two other structures is  
242 that the  $\beta$ -helix of Calhy\_0908C is three turns longer than that of Calkro\_0844C and  
243 Calkr\_0826C, which is made possible by a massive 62 residue single insert in the Calhy\_0908C  
244 sequence. The rest of the Calhy\_0908C (residues 170-378 and 443-578) could be  
245 superimposed reasonably well onto Calkro\_0844C with the r.m.s.d. of 1.97 Å over 1,020 atoms.

246 Aside from additional turns of the  $\beta$ -helix, Calhy\_0908C has other differences from  
247 Calkro\_0844C (**Figure 5B**). Similar to the Calkr\_0826C vs. Calkro\_0844C contrast (**Figure 5A**),  
248 the set of loops connecting faces B and C of the  $\beta$ -helix is different. The loop connecting  $\beta$ 5 and  
249  $\beta$ 6 is 7 residues longer in Calhy\_0908C and four more neighboring loops are shorter in  
250 Calhy\_0908C: the  $\beta$ 8- $\beta$ 9 loop,  $\beta$ 11- $\beta$ 12 loop,  $\beta$ 39- $\beta$ 40 loop, and  $\beta$ 42- $\beta$ 43 loop is 7, 2, 2, and 5  
251 residues shorter, respectively. Another difference is the orientation of the helix  $\alpha$ 3, which while  
252 still present in the Calhy\_0908C structure, runs almost perpendicular to the corresponding helix

253 of Calkro\_0844C. To complement that rearrangement, the connecting loop in Calhy\_0908C  
254 goes straight to the  $\beta$ 36 without an 8-residue 'detour' that is present in Calkro\_0844C. Also, the  
255  $\beta$ 29 strand of Calkro\_0844C is not present in Calhy\_0908C, with its place taken by the  
256 repositioned  $\alpha$ 3 helix. Similar to both Calkro\_0844C and Calkr\_0826C, in the hydrophobic face  
257 A of Calhy\_0908C, the  $\beta$ -helix features a line of aromatic residues (**Figure 6A-C**).

258 **Can the observed differences in tāpirin structures explain the differences in**  
259 **cellulose binding?** As shown in the **Figure 4**, Calhy\_0908 binds to more sites on cellulose  
260 compared to the other tāpirins examined here. When the structural features of three of these  
261 tāpirins are compared (**Figure 5C**), there are two regions where differences are apparent. First,  
262 the set of loops, connecting  $\beta$ -strands on the edge between faces B and C of the  $\beta$ -helix, is  
263 different in these three proteins, varying in size and chemistry. However, it should be pointed  
264 out that Calhy\_0908C has the least extensive set of loops, such that most of these are shorter  
265 than the corresponding regions in Calkr\_0826C and Calkro\_0844C. If these loops were  
266 responsible for the protein interaction with the cellulose, this difference would leave overall less  
267 exposed surface area for the possible protein-cellulose interactions. However, that cannot be  
268 the case, as Calhy\_0908 does indeed seem to bind well to cellulose despite the differences in  
269 the loop sets.

270 Still, another area of interest is the hydrophobic surface found on the face A of the  $\beta$ -  
271 helix that is protected by the connecting loop in the crystallized conformations. We again tried  
272 observe an interaction between the tāpirins and soluble cellooligosaccharides (C2 up to C6),  
273 however similar to Calkro\_0844C (21), neither Calkr\_0826C or Calhy\_0908C co-crystallized or  
274 interacted with the oligosaccharides. Regardless, a possible cellulose-binding mechanism could  
275 involve repositioning of the connecting loop upon mechanical contact with the cellulose surface,  
276 exposing the line of aromatic sidechains positioned flat on the surface and spaced 5 Å apart,  
277 which corresponds to the distance between sugar units in the cellulose chain. This can be seen

278 in **Figure 6**, where Calhy\_0908C has the largest hydrophobic surface of the three tāpirins as  
279 well as the largest number of aromatic side chains lined up on that surface.

280

281 **Localization of tāpirins on the cell surface of *Caldicellulosiruptor*.** There is  
282 evidence for tāpirin localization on the cell surface, based on immunofluorescence microscopy  
283 using antibodies directed against these proteins (**Figure 7**). Tāpirins are most evident at the  
284 poles of the cell, although they also appear to decorate the cellular surface. Given their  
285 proposed role as cellulose-binding proteins, especially for initial attachment to substrate, this is  
286 consistent with that hypothesis.

287 **Summary.** It interesting that binding assays to cellulosic substrates indicated that  
288 Calhy\_0908 and Calkr\_0826 from the weakly cellulolytic species *C. hydrothermalis* and *C.*  
289 *kristjanssonii*, respectively, appeared to bind in higher quantities to cellulose than those tāpirins  
290 from prolific microcrystalline cellulose degraders. Structures from the C-termini of both  
291 Calhy\_0908 and Calkr\_0826, compared to the previously characterized Calkro\_0844 (21),  
292 identified clear differences between the tāpirins, including a much longer potential binding  
293 platform in Calhy\_0908, which contains the largest number of hydrophobic and aromatic  
294 residues of the three. Whether the density of tāpirins on the cell surface varies across  
295 *Caldicellulosiruptor* species is not yet clear. But if somewhat equivalent within the genus, *C.*  
296 *hydrothermalis* may use tāpirin-based adhesion as mechanism to support survival of a less  
297 cellulolytic species in thermal environments.

298 While fluorescence microscopy indicates that the presence of tāpirins on the outside of  
299 the *Caldicellulosiruptor* cells, their exact cellular location needs to be resolved, especially as this  
300 relates to possible association with Type IV pili. Another interesting question going forward is  
301 whether tāpirins are a uniquely *Caldicellulosiruptor* feature or whether counterparts are used by  
302 other microorganisms to attach to substrates or surfaces.

303

304

## Materials and Methods

305

306

307

**Bacterial strains, plasmids, and substrates.** The wild type strain of  
308 *Caldicellulosiruptor bescii* was obtained from Leibniz Institute DSMZ-German Collection of  
309 Microorganisms and Cell Cultures, and *C. bescii* strain MACB1018 and genetic vector,  
310 pGL0100, were developed previously (27). *Escherichia coli* strains 5-alpha (New England  
311 BioLabs) and Rosetta (Millipore Sigma, Merck) were used for plasmid replication and protein  
312 production, respectively. Genes of interest were PCR amplified from extracted genomic DNA, as  
313 described previously (28), and Gibson assembly ((29) Gibson assembly master mix, New  
314 England BioLabs) or a KLD reaction (KLD Enzyme Mix, New England BioLabs) was used to  
315 insert the fragments into plasmids, which had been extracted with ZymoPure midiprep and  
316 Zymo Research plasmid miniprep classic kits (Zymo Research). Additionally, Athe\_1870  
317 (GenBank™ accession number WP\_015908253.1), Calhy\_0908 (GenBank™ accession  
318 number YP\_003992006.1), Calkr\_0826 (GenBank™ accession number YP\_004025962.1),  
319 NA10\_0869 (DOE Joint Genome Institute (JGI) Integrated Microbial Genomes & Microbiomes  
320 gene ID 2566072544), and Calkro\_0844 (GenBank™ accession number YP\_004023543.1)  
321 genes were *E. coli* codon optimized (without transmembrane domains or signal peptides) with  
322 Integrated DNA Technologies (IDT) Codon Optimization Tool  
323 (<https://www.idtdna.com/CodonOpt>), and synthesized by the DOE JGI on a pET-45 plasmid. All  
324 proteins produced included an *N*-terminal hexahistidine affinity tag. Sequences of all plasmids  
325 and edited portions of final strains were confirmed with Sanger sequencing (Genewiz).  
326 Substrates used for growth and binding include: Avicel PH-101 (FMC BioPolymer), Cave-in-  
327 Rock switchgrass (*Panicum virgatum* L. from fields in Monroe County, IA, retrieved by the  
328 National Renewable Energy Laboratory, and ground and sieved using a Wiley mill (Thomas

329 Scientific) and 40/80 mesh, respectively), beechwood xylan (Sigma-Aldrich), and poplar  
330 (*Populus trichocarpa*, obtained from Vincent Chiang (30))

331  
332 **Production of tāpirin proteins in *E. coli*.** Expression plasmids (pET-45, see above)  
333 with synthesized Athe\_1870, Calhy\_0908, Calkr\_0826, NA10\_0869, and Calkro\_0844 genes  
334 were transformed into *E. coli* (strain 5-alpha with 50 µg/ml carbenicillin selection and strain  
335 Rosetta with both 50 µg/ml carbenicillin and 34 µg/ml chloramphenicol selection), and cultured  
336 on Luria-Bertani (LB – 10 g/liter sodium chloride, 10 g/liter tryptone, and 5 g/L yeast extract)  
337 liquid medium or agar (1.5% wt/vol) plates at 37°C. For the production of protein, the cultures  
338 were grown in ZYM-5052 autoinduction medium (31) with chloramphenicol and carbenicillin in  
339 up to 1-3L volumes at 37°C 250 RPM for 18-24 hours and harvested with centrifugation of 6000  
340 x g for 10 minutes. Cells were then re-suspended in 100 mL of 20 mM sodium phosphate, pH  
341 7.4, 0.5 mM sodium chloride, and 5 mM imidazole, lysed with a French Press at 16,000 psig,  
342 heat-treated at 65°C for 30 minutes, and centrifuged at 25,000 x g for 30 minutes. Protein in  
343 soluble fraction was purified with 5-ml HisTrap HP nickel-Sepharose immobilized metal affinity  
344 chromatography column (GE Healthcare - operated according to the manufacturer's  
345 instructions) using a Biologic DuoFlow FPLC (Bio-Rad), and then stored at 4°C. Protein  
346 concentration was determined by the Bradford assay (32) and protein purity was visualized  
347 along with a Benchmark protein ladder (Life Technologies) by SDS-PAGE using 4-15% Mini-  
348 PROTEAN® TGX Stain-Free™ Precast Gels (Bio-Rad).

349 **Production and purification of C-terminal tāpirins, Calkr\_0826C and Calhy\_0908C.**  
350 Purified protein was buffer-exchanged into 0.4 mM calcium chloride, 0.15 M sodium chloride,  
351 and 50 mM tris-chloride (pH 8.0) reaction buffer and treated with thermolysin (1 mg/ml in same  
352 buffer - Promega), as described in (21). A thermolysin-to-protein ratio of 1:500 and treatment  
353 times of 5 to 30 minutes at 70°C were used to effectively lyse protein to produce the C-terminal  
354 portion of the tāpirin (i.e. Calkr\_0826C for *C. kristjanssonii* and Calhy\_0809C for *C.*

355 *hydrothermalis*). Reactions were halted by storing samples on ice and then long term at 4°C.  
356 Samples were imaged with SDS-PAGE as described above to verify cleavage. For the  
357 crystallization of Calkr\_0826C and Calhy\_0908C, cleaved products were further purified using  
358 an ÄKTA protein purification system (GE Healthcare Life Sciences) and Superdex 75 pg (16/60)  
359 size exclusion chromatography column in 20 mM Tris pH 7.5 and 100 mM sodium chloride.

360 **Crystallization.** The crystals of Calkr\_0826C and Calhy\_0908C were initially obtained  
361 with sitting drop vapor diffusion using a 96-well plate with PEG ion HT screen from Hampton  
362 Research (Aliso Viejo, CA). 50 µL of well solution was added to the reservoir, and drops were  
363 made with 0.2 µL of well solution and 0.2 µL of protein solution using a Phoenix crystallization  
364 robot (Art Robbins Instruments, Sunnyvale, CA). The Calkr\_0826C crystals were grown at 20°C  
365 using an optimization screen containing 0.1 M citric acid pH 3.0 to 4.0 and 10% to 15% w/v  
366 polyethylene glycol (PEG) 3350 (best crystals appeared in pH range from 3.1 to 3.2 and PEG  
367 3350 14-15%). The protein solutions contained 12 mg/mL of protein, 20 mM Tris pH 7.5, 100  
368 mM sodium chloride, 2% of the Hampton Research Tacsimate pH 4 mix and 5 mM of each of  
369 zinc acetate, potassium chloride, magnesium chloride, and calcium chloride. The Calhy\_0908C  
370 crystals were grown at 20°C using an optimization screen containing 5 mM – 35 mM zinc  
371 acetate and 15% to 24% w/v PEG 3350 (best crystals appeared in 0.015 M zinc acetate and  
372 PEG 3350 concentration of 17-18%). The protein solutions contained 7.5 mg/mL of protein, 20  
373 mM Tris pH 7.5, 100 mM NaCl and 2% of the Hampton Research Tacsimate pH 7 mix.

374 All crystals were soaked in well solution with PEG 3350 increased to 25% along with 5-  
375 10% ethylene glycol added for the cryo protection. For the purpose of structure determination,  
376 an iodine derivative was obtained for the Calhy\_0908C by quick soaking the crystals in the  
377 cryoprotectant described above with 0.1 M potassium iodide added.

378 **Crystallography data collection and processing.** The crystals of Calkr\_0826C and  
379 Calhy\_0908C were flash frozen in a nitrogen gas stream at 100 K before home source data

380 collection using an in-house Bruker X8 MicroStar X-Ray generator with Helios mirrors and  
381 Bruker Platinum 135 CCD detector. Data were indexed and processed with the Bruker Suite of  
382 programs version 2014.9 (Bruker AXS, Madison, WI).

383 **Crystal structure solution and refinement.** Intensities were converted into structure  
384 factors and 'free' sets of the reflections (5% of the reflections for Calkr\_0826C and 2% for  
385 Calhy\_0908C) were flagged for  $R_{\text{free}}$  calculations using programs F2MTZ, Truncate, CAD and  
386 Unique from the CCP4 package of programs (33). The structure of the Calkr\_0826C was solved  
387 by MOLREP (34) using Calkro\_0844\_C (21) (PDB ID 4WA0) as a search model. Crank2 (35)  
388 was used to solve the structure of the Calhy\_0908C utilizing iodine single-wavelength  
389 anomalous dispersion (36). Refinement and manual correction was performed using REFMAC5  
390 (37) version 5.8.158, PHENIX (38) version 1.11 and Coot (39) version 0.8.8. The  
391 MOLPROBITY method (40) was used to analyze the Ramachandran plot, and root mean  
392 square deviations (rmsd) of bond lengths and angles were calculated from ideal values of Engh  
393 and Huber stereo chemical parameters (41). Wilson B-factor was calculated using  
394 CTRUNCATE version 1.15.10 (33). The data collection and refinement statistics are shown in  
395 **Table S3**. Structures of the tapirins' C-terminal domains were deposited to the RCSB PDB with  
396 the access codes 6N2B for the Calkr\_0826 and 6N2C for the Calhy\_0908.

397

398 **Tāpirin binding assays.** Recombinant proteins, both full length and truncated, were  
399 tested for attachment to various substrates in triplicate, as described in (21). All substrates were  
400 initially soaked with 100 mL of 50 mM MES and 3.9 mM sodium chloride at pH 7.2 ('binding  
401 buffer') overnight, and then subsequently dried overnight (both steps at 70°C). Nine mg of  
402 washed substrates were mixed with 40 µg of purified tāpirin protein and incubated in a  
403 thermomixer (Eppendorf) at 70°C and 500 rpm for one hour. Samples were then centrifuged  
404 13000 x g and separated into 'unbound' (supernatant) and 'bound' (pellet) fractions. The bound

405 fraction was then washed (re-suspending the substrate in binding buffer with vortexing,  
406 centrifuging mixture at 13000 x g, and discarding the supernatant) four times before being finally  
407 re-suspended in 250  $\mu$ L of buffer prior to SDS-PAGE. Equal volumes of bound or unbound  
408 sample were mixed with 2x Laemmli sample buffer and 5% 2-mercaptoethanol, and boiled for  
409 30 minutes. Samples were then loaded on a SDS-PAGE gel as described above. Densitometry  
410 was completed by using ImageJ (42) to analyze band intensity (keeping contrast across gels  
411 consist) and by normalizing all bands to the 70 kDa Benchmark protein ladder (Life  
412 Technologies) band present on each gel.

413  
414 **Deletion of *tāpirin* and/or *pili* genes in *C. bescii*.** Knockout (KO) vectors were  
415 constructed with Gibson Assembly with pGL100 (27) as the backbone with *pyrE* (Athe\_1382).  
416 Flanking regions outside Athe\_1870-1871, and Athe\_1870-1885 were PCR-amplified using *C.*  
417 *bescii* MACB1018 genomic DNA as a template, while the vector backbone and kanamycin  
418 resistance gene (HTK) and SLP promoter (Athe\_2303) and were PCR-amplified from a template  
419 plasmid before using Gibson Assembly to assemble the knock out vectors (see below). A new  
420 genetic vector, pLLL023, was also generated using a KLD reaction from pGL100 to create a  
421 plasmid without P<sub>slp</sub>-HTK on the backbone, such that P<sub>slp</sub>-HTK could be inserted into the  
422 genome. Primers used for vector construction and genetic screening are included in **Table 2**.  
423 The resulting vectors were generated for the following strains: pLLL024 for RKCB136  
424 (Athe\_1870-1871 KO and P<sub>slp</sub>-HTK knock- in (KI)), and pLLL012 for RKCB135 (Athe\_1870-  
425 1885 KO). After construction, plasmids were methylated with purified recombinant M.CbeI (as  
426 described previously in (27, 43)).

427  
428 *C. bescii* genetic strains were all cultured anaerobically in either low osmolality defined  
429 (LOD) or complex (LOC) medium (44) with cellobiose in liquid cultures at 75°C with a nitrogen  
430 headspace and on 1.5% wt/vol agar plates in an anaerobic chamber (Coy Laboratory) at 65°C.  
431 To transform KO plasmids into *C. bescii* strain MACB1018 (27), 1  $\mu$ g of plasmid DNA was

432 added to 50  $\mu$ l aliquots of competent cells (prepped as described previously (27) in LOD media  
433 supplemented with 1 X 19 amino acid solution (45)) at room temperature. Cells were then  
434 electroporated with a Gene Pulser II system with a Pulse Controller Plus module (Bio-Rad) in 1-  
435 mm-gap cuvettes (USA Scientific) at 25  $\mu$ F, 200  $\Omega$ , and 2 kV before being passaged to 10 mL of  
436 preheated LOC medium for recovery at 75°C. After one hour, all recovery media was passaged  
437 to selective media (LOC with 50  $\mu$ g/mL kanamycin) and incubated until growth was noted  
438 (typically 1 to 3 days) at 75°C. Growing transformants were screened with PCR and then  
439 passaged multiple times into liquid selective medium and/or plated on selective media for  
440 purification. Second crossovers were then selected by plating transformed cells onto LOD  
441 medium containing 4 mM 5-fluoroorotic acid (5-FOA), 50  $\mu$ g/mL kanamycin (for HTK KI strains),  
442 and 40 mM uracil. Successful PCR-screened colonies were plate purified on the same second-  
443 crossover plating media without 5-FOA.

444  
445 **Whole cell binding assay.** *Caldicellulosiruptor* cells were cultured at 75°C in 100 mL of  
446 modified 671 medium with 5 g/L Avicel, as described previously. Cultures were measured for  
447 cell density with acridine orange epifluorescence microscopy, as described previously (46),  
448 before the entire sample was initially centrifuged at 400 x g for 5 minutes to lightly pellet Avicel.  
449 The supernatant was removed and cells were harvested by centrifugation at 6000 x g for 10  
450 minutes before being concentrated to 5x10<sup>8</sup> to 1x10<sup>9</sup> cells/ml in 671d medium. 1 mL of cells  
451 were added to 10 mg of Avicel for 1 hour in a thermomixer (Eppendorf) at 70°C; a control was  
452 also completed without any substrate (only media and cells) and all samples were completed in  
453 triplicate. The supernatant counting unbound cells were then separated and enumerated again  
454 with acridine orange epifluorescence microscopy, with the average planktonic cell densities  
455 were compared using a t-test.

456  
457 **Acknowledgements**

458  
459 This work was supported by the BioEnergy Science Center (BESC), a U.S. Department

460 of Energy Bioenergy Research Center supported by the Office of Biological and Environmental  
461 Research in the DOE Office of Science. LL Lee acknowledges support from a National Science  
462 Foundation Graduate Research Fellowship and a NIH T32 Biotechnology Traineeship  
463 (GM008776-11). Synthetic genes for the production of the tāpirins were provided by the U.S.  
464 Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, which is  
465 supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-  
466 AC02-05CH11231.  
467

468  
469**References**

- 470 1. Blumer-Schuetz S, Kataeva I, Westpheling J, Adams M, Kelly R. 2008. Extremely  
471 thermophilic microorganisms for biomass conversion: status and prospects. *Curr Opin*  
472 *Biotechnol* 19:210-217.
- 473 2. Blumer-Schuetz S, Giannone R, Zurawski J, Ozdemir I, Ma Q, Yin Y, Xu Y, Kataeva I,  
474 Poole F, Adams M, Hamilton-Brehm S, Elkins J, Larimer F, Land M, Hauser L,  
475 Cottingham R, Hettich R, Kelly R. 2012. *Caldicellulosiruptor* core and pangenomes  
476 reveal determinants for noncellulosomal thermophilic deconstruction of plant biomass. *J*  
477 *Bacteriol* 194:4015-4028.
- 478 3. Lee L, Blumer-Schuetz S, Izquierdo J, Zurawski J, Loder A, Conway J, Elkins J, Podar  
479 M, Clum A, Jones P, Piatek M, Weighill D, Jacobson D, Adams M, Kelly R. 2018.  
480 Genus-wide assessment of lignocellulose utilization in the extremely thermophilic  
481 *Caldicellulosiruptor* by genomic, pan-genomic and metagenomic analysis. *Appl Environ*  
482 *Microbiol* doi:10.1128/aem.02694-17.
- 483 4. Conway J, Zurawski J, Lee L, Blumer-Schuetz S, Kelly R. 2015. Lignocellulosic  
484 biomass deconstruction by the extremely thermophilic genus *Caldicellulosiruptor*  
485 doi:10.21775/9781910190135.04. Caister Academic Press, Norfolk, UK.
- 486 5. Zurawski J, Blumer-Schuetz S, Conway J, Kelly R. 2014. The extremely thermophilic  
487 genus *Caldicellulosiruptor*: physiological and genomic characteristics for complex  
488 carbohydrate conversion to molecular hydrogen, p 177-195. *In* Zannoni D, De Philippis  
489 R (ed), *Microbial BioEnergy: Hydrogen Production* doi:10.1007/978-94-017-8554-9\_8.  
490 Springer Netherlands, Dordrecht.
- 491 6. Bayer E, Lamed R, White B, Flint H. 2008. From cellulosomes to cellulosomes. *Chem*  
492 *Rec* 8:364-377.
- 493 7. van den Brink J, de Vries R. 2011. Fungal enzyme sets for plant polysaccharide

- 494 degradation. Appl Environ Microbiol 91:1477.
- 495 8. Lombard V, Golaconda Ramulu H, Drula E, Coutinho P, Henrissat B. 2014. The  
496 carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490-  
497 D495.
- 498 9. Brunecky R, Alahuhta M, Xu Q, Donohoe B, Crowley M, Kataeva I, Yang S, Resch M,  
499 Adams M, Lunin V, Himmel M, Bomble Y. 2013. Revealing nature's cellulase diversity:  
500 the digestion mechanism of *Caldicellulosiruptor bescii* CelA. Science 342:1513-6.
- 501 10. Yi Z, Su X, Revindran V, Mackie RI, Cann I. 2013. Molecular and biochemical analyses  
502 of CbCel9A/Cel48A, a highly secreted multi-modular cellulase by *Caldicellulosiruptor*  
503 *bescii* during growth on crystalline cellulose. PLoS One 8:e84172.
- 504 11. Brunecky R, Donohoe B, Yarbrough J, Mittal A, Scott B, Ding H, Taylor li L, Russell J,  
505 Chung D, Westpheling J, Teter S, Himmel M, Bomble Y. 2017. The multi domain  
506 *Caldicellulosiruptor bescii* CelA cellulase excels at the hydrolysis of crystalline cellulose.  
507 Sci Rep 7:9622.
- 508 12. Chung D, Young J, Bomble Y, Vander Wall T, Groom J, Himmel M, Westpheling J.  
509 2015. Homologous expression of the *Caldicellulosiruptor bescii* CelA reveals that the  
510 extracellular protein is glycosylated. PLoS One 10:e0119508.
- 511 13. Kim S, Chung D, Himmel M, Bomble Y, Westpheling J. 2017. Engineering the N-terminal  
512 end of CelA results in improved performance and growth of *Caldicellulosiruptor bescii* on  
513 crystalline cellulose. Biotechnol Bioeng 114:945-950.
- 514 14. Young J, Chung D, Bomble Y, Himmel M, Westpheling J. 2014. Deletion of  
515 *Caldicellulosiruptor bescii* CelA reveals its crucial role in the deconstruction of  
516 lignocellulosic biomass. Biotechnol Biofuels 7:142.
- 517 15. Zverlov V, Mahr S, Riedel K, Bronnenmeier K. 1998. Properties and gene structure of a  
518 bifunctional cellulolytic enzyme (CelA) from the extreme thermophile '*Anaerocellum*  
519 *thermophilum*' with separate glycosyl hydrolase family 9 and 48 catalytic domains.

- 520 Microbiology 144 ( Pt 2):457-65.
- 521 16. Xue X, Wang R, Tu T, Shi P, Ma R, Luo H, Yao B, Su X. 2015. The n-terminal GH10  
522 domain of a multimodular protein from *Caldicellulosiruptor bescii* is a versatile  
523 xylanase/beta-glucanase that can degrade crystalline cellulose. Appl Environ Microbiol  
524 81:3823-33.
- 525 17. Meng D, Ying Y, Chen X, Lu M, Ning K, Wang L, Li F. 2015. Distinct roles for  
526 carbohydrate-binding modules of glycoside hydrolase 10 (GH10) and GH11 xylanases  
527 from *Caldicellulosiruptor* sp. strain F32 in thermostability and catalytic efficiency. Appl  
528 Environ Microbiol 81:2006-14.
- 529 18. Conway J, Pierce W, Le J, Harper G, Wright J, Tucker A, Zurawski J, Lee L, Blumer-  
530 Schuette S, Kelly R. 2016. Multi-Domain, surface layer associated glycoside hydrolases  
531 contribute to plant polysaccharide degradation by *Caldicellulosiruptor* species. J Biol  
532 Chem 291:6732-6747.
- 533 19. Ozdemir I, Blumer-Schuette S, Kelly R. 2012. S-layer homology domain proteins  
534 Csac\_0678 and Csac\_2722 are implicated in plant polysaccharide deconstruction by the  
535 extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. Appl Environ  
536 Microbiol 78:768-77.
- 537 20. Yokoyama H, Yamashita T, Morioka R, Ohmori H. 2014. Extracellular secretion of  
538 noncatalytic plant cell wall-binding proteins by the cellulolytic thermophile  
539 *Caldicellulosiruptor bescii*. J Bacteriol 196:3784-92.
- 540 21. Blumer-Schuette S, Alahuhta M, Conway J, Lee L, Zurawski J, Giannone R, Hettich R,  
541 Lunin V, Himmel M, Kelly R. 2015. Discrete and structurally unique proteins (täpirins)  
542 mediate attachment of extremely thermophilic *Caldicellulosiruptor* species to cellulose. J  
543 Biol Chem 290:10645-10656.
- 544 22. Conway J, McKinley B, Seals N, Hernandez D, Khatibi P, Poudel S, Giannone R, Hettich  
545 R, Williams-Rhaesa A, Lipscomb G, Adams M, Kelly R. 2017. Functional analysis of the

- 546 Glucan Degradation Locus (GDL) in *Caldicellulosiruptor bescii* reveals essential roles of  
547 component glycoside hydrolases in plant biomass deconstruction. Appl Environ  
548 Microbiol doi:10.1128/aem.01828-17.
- 549 23. Park S, Johnson D, Ishizawa C, Parilla P, Davis M. 2009. Measuring the crystallinity  
550 index of cellulose by solid state <sup>13</sup>C nuclear magnetic resonance. Cellulose 16:641-647.
- 551 24. Hong J, Ye X, Zhang Y. 2007. Quantitative determination of cellulose accessibility to  
552 cellulase based on adsorption of a nonhydrolytic fusion protein containing CBM and GFP  
553 with its applications. Langmuir 23:12535-12540.
- 554 25. Lochner A, Giannone R, Keller M, Antranikian G, Graham D, Hettich R. 2011. Label-free  
555 quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor*  
556 *obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline  
557 cellulose, and switchgrass. J Proteome Res 10:5302-5314.
- 558 26. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T. 2008.  
559 NCBI/BLAST: a better web interface. Nucleic Acids Res 36:W5-W9.
- 560 27. Lipscomb G, Conway J, Blumer-Schuetz S, Kelly R, Adams M. 2016. A highly  
561 thermostable kanamycin resistance marker expands the tool kit for genetic manipulation  
562 of *Caldicellulosiruptor bescii*. Appl Environ Microbiol 82:4421-4428.
- 563 28. Geslin C, Le Romancer M, Erauso G, Gaillard M, Perrot G, Prieur D. 2003. PAV1, the  
564 first virus-like particle isolated from a hyperthermophilic euryarchaeote, "*Pyrococcus*  
565 *abyssi*". J Bacteriol 185:3888-3894.
- 566 29. Gibson D, Young L, Chuang R, Venter J, Hutchison Iii A, Smith HO. 2009. Enzymatic  
567 assembly of DNA molecules up to several hundred kilobases. Nature Methods 6:343.
- 568 30. Wang J, Matthews M, Williams C, Shi R, Yang C, Tunlaya-Anukit S, Chen H, Li Q, Liu J,  
569 Lin C, Naik P, Sun Y, Loziuk P, Yeh T, Kim H, Gjersing E, Shollenberger T, Shuford C,  
570 Song J, Miller Z, Huang Y, Edmunds C, Liu B, Sun Y, Lin Y, Li W, Chen H, Peszlen I,  
571 Ducoste J, Ralph J, Chang H, Muddiman D, Davis M, Smith C, Isik F, Sederoff R,

- 572 Chiang V. 2018. Improving wood properties for wood utilization through multi-omics  
573 integration in lignin biosynthesis. *Nat Commun* 9:1579.
- 574 31. Studier F. 2005. Protein production by auto-induction in high density shaking cultures.  
575 *Protein Expr Purif* 41:207-34.
- 576 32. Bradford M. 1976. A rapid and sensitive method for the quantitation of microgram  
577 quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72:248-  
578 254.
- 579 33. Winn M, Ballard C, Cowtan K, Dodson E, Emsley P, Evans P, Keegan R, Krissinel E,  
580 Leslie A, McCoy A, McNicholas S, Murshudov G, Pannu N, Potterton E, Powell H, Read  
581 R, Vagin A, Wilson K. 2011. Overview of the CCP4 suite and current developments.  
582 *Acta Crystallogr D Biol Crystallogr* 67:235-42.
- 583 34. Vagin A, Teplyakov A. 2010. Molecular replacement with MOLREP. *Acta Crystallogr D*  
584 *Biol Crystallogr* 66:22-5.
- 585 35. Skubak P, Pannu N. 2013. Automatic protein structure solution from weak x-ray data.  
586 *Nat Commun* 4.
- 587 36. Hendrickson W, Teeter M. 1981. Structure of the hydrophobic protein crambin  
588 determined directly from the anomalous scattering of sulphur. *Nature* 290:107-113.
- 589 37. Murshudov G, Skubak P, Lebedev A, Pannu N, Steiner R, Nicholls R, Winn M, Long F,  
590 Vagin A. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta*  
591 *Crystallogr D Biol Crystallogr* 67:355-67.
- 592 38. Afonine P, Grosse-Kunstleve R, Echols N, Headd J, Moriarty N, Mustyakimov M,  
593 Terwilliger T, Urzhumtsev A, Zwart P, Adams P. 2012. Towards automated  
594 crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol*  
595 *Crystallogr* 68:352-67.
- 596 39. Emsley P, Lohkamp B, Scott W, Cowtan K. 2010. Features and development of Coot.  
597 *Acta Crystallogr D Biol Crystallogr* 66:486-501.

- 598 40. Chen VB, Arendall W, 3rd, Headd J, Keedy D, Immormino R, Kapral G, Murray L,  
599 Richardson J, Richardson D. 2010. MolProbity: all-atom structure validation for  
600 macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12-21.
- 601 41. Engh R, Huber R. 1991. Accurate bond and angle parameters for x-ray protein-structure  
602 refinement. *Acta Crystallographica Section A* 47:392-400.
- 603 42. Schneider C, Rasband W, Eliceiri K. 2012. NIH Image to ImageJ: 25 years of image  
604 analysis. *Nat Methods* 9:671-5.
- 605 43. Chung D, Huddleston J, Farkas J, Westpheling J. 2011. Identification and  
606 characterization of Cbel, a novel thermostable restriction enzyme from  
607 *Caldicellulosiruptor bescii* DSM 6725 and a member of a new subfamily of HaeIII-like  
608 enzymes. *J Ind Microbiol Biotechnol* 38:1867-77.
- 609 44. Farkas J, Chung D, Cha M, Copeland J, Grayeski P, Westpheling J. 2013. Improved  
610 growth media and culture techniques for genetic analysis and assessment of biomass  
611 utilization by *Caldicellulosiruptor bescii*. *J Ind Microbiol Biotechnol* 40:41-49.
- 612 45. Lipscomb G, Stirrett K, Schut G, Yang F, Jenney F, Scott R, Adams M, Westpheling J.  
613 2011. Natural competence in the hyperthermophilic archaeon *Pyrococcus furiosus*  
614 facilitates genetic manipulation: construction of markerless deletions of genes encoding  
615 the two cytoplasmic hydrogenases. *Appl Environ Microbiol* 77:2232-2238.
- 616 46. VanFossen A, Ozdemir I, Zelin S, Kelly R. 2011. Glycoside hydrolase inventory drives  
617 plant polysaccharide deconstruction by the extremely thermophilic bacterium  
618 *Caldicellulosiruptor saccharolyticus*. *Biotechnol Bioeng* 108:1559-1569.
- 619  
620

621 **FIGURE CAPTIONS**

622 **Figure 1. Genomic organization of Type IV pili, tāpirins, and glucan degradation loci in**  
623 **examined *Caldicellulosiruptor* species.** Numbers refer to gene locus tags in sectioned loci.  
624 The tāpirins that were examined in this study are highlighted in purple. Abbreviations are as  
625 follows: Athe, *C. bescii*; Calhy, *C. hydrothermalis*; Calkr, *C. kristjanssonii*; Calkro, *C.*  
626 *kronotskyensis*; NA10, *C. naganoensis*.

627

628 **Figure 2. Whole cell binding assay of parent *Caldicellulosiruptor bescii* strain,**  
629 **MACB1018, versus *C. bescii* tāpirin and pili deletion ( $\Delta$ Athe\_1870-1885 – RKCB135) and**  
630 **tāpirin only deletion ( $\Delta$ Athe\_1870-1871 – RKCB136) strains.** Cell concentrations refer to  
631 cells not attached to Avicel and/or test tube wall after 1 h incubation. Binding assays included  
632  $5 \times 10^8$  to  $1 \times 10^9$  cells/ml in 671d medium incubated with or without 10 mg of Avicel for 1 hour at  
633 70°C. Planktonic cell concentrations were quantified using epifluorescent microscopy. Error bars  
634 are standard error of triplicate samples (\* indicates statistical significance; 'n.s.' is 'not  
635 significant').

636

637 **Figure 3. SDS-PAGE gel of tāpirin binding assay to plant component substrates.**  
638 Recombinant Athe\_1870, Calhy\_0908, Calkr\_0826, Calkro\_0844, and NA10\_0869 cloned and  
639 produced from *C. hydrothermalis*, *C. kristjanssonii*, *C. kronotskyensis*, and *C. naganoensis*,  
640 respectively, were incubated with cellulose (Avicel and filter paper), lignocellulose (switchgrass  
641 and poplar), and xylan, along with a no substrate control. Proteins were separated into bound  
642 (B) and unbound (U) fractions and visualized with SDS-PAGE; bands here are representative of  
643 triplicate trials.

644

645 **Figure 4. Densitometry of recombinant proteins bound to cellulosic substrates as**  
646 **visualized with SDS-PAGE.** Bound bands of recombinant Athe\_1870 (gray), Calhy\_0908

647 (orange), Calkr\_0826 (yellow), Calkro\_0844 (light blue), and NA10\_0869 (dark blue) from *C.*  
648 *bescii*, *C. hydrothermalis*, *C. kristjanssonii*, *C. kronotskyensis*, and *C. naganoensis*, respectively,  
649 were quantified and normalized by the 70 kDa Benchmark ladder band on associated SDS-  
650 PAGE gel. Error bars represent the standard deviations of the intensities of triplicate samples.

651

652 **Figure 5. Crystal structures of Calkr\_0826, Calhy\_0908, and Calkro\_0844.** A) Calkr\_0826  
653 and Calkro\_0844 superimposed; B) Calhy\_0908 and Calkr\_0844 superimposed; C)  
654 Calhy\_0908, Calkr\_0826, and Calkro\_0844 superimposed; and D) 90° rotation of Calhy\_0908,  
655 showing  $\beta$ -helix face designations: 'A', 'B', and 'C'. The colors refer to the following tāpirins:  
656 green, Calhy\_0908 (*C. hydrothermalis*); blue, Calkr\_0826 (*C. kristjanssonii*); and magenta,  
657 Calkro\_0844 (*C. kronotskyensis*). Abbreviations are as follows:  $\alpha$ , alpha helix; N, N-terminus;  
658 and C, C-terminus. Colors of labels correspond to the colors of the peptides; black font refers to  
659 a shared feature across all tāpirins.

660

661 **Figure 6. Surface features of tāpirins: A) Calhy\_0908 (*C. hydrothermalis*), B) Calkro\_0844**  
662 **(*C. kronotskyensis*), and C) Calkr\_0826 (*C. kristjanssonii*).** Connecting loop and N- and C-  
663 termini are removed. Aromatic sidechains are highlighted in red.

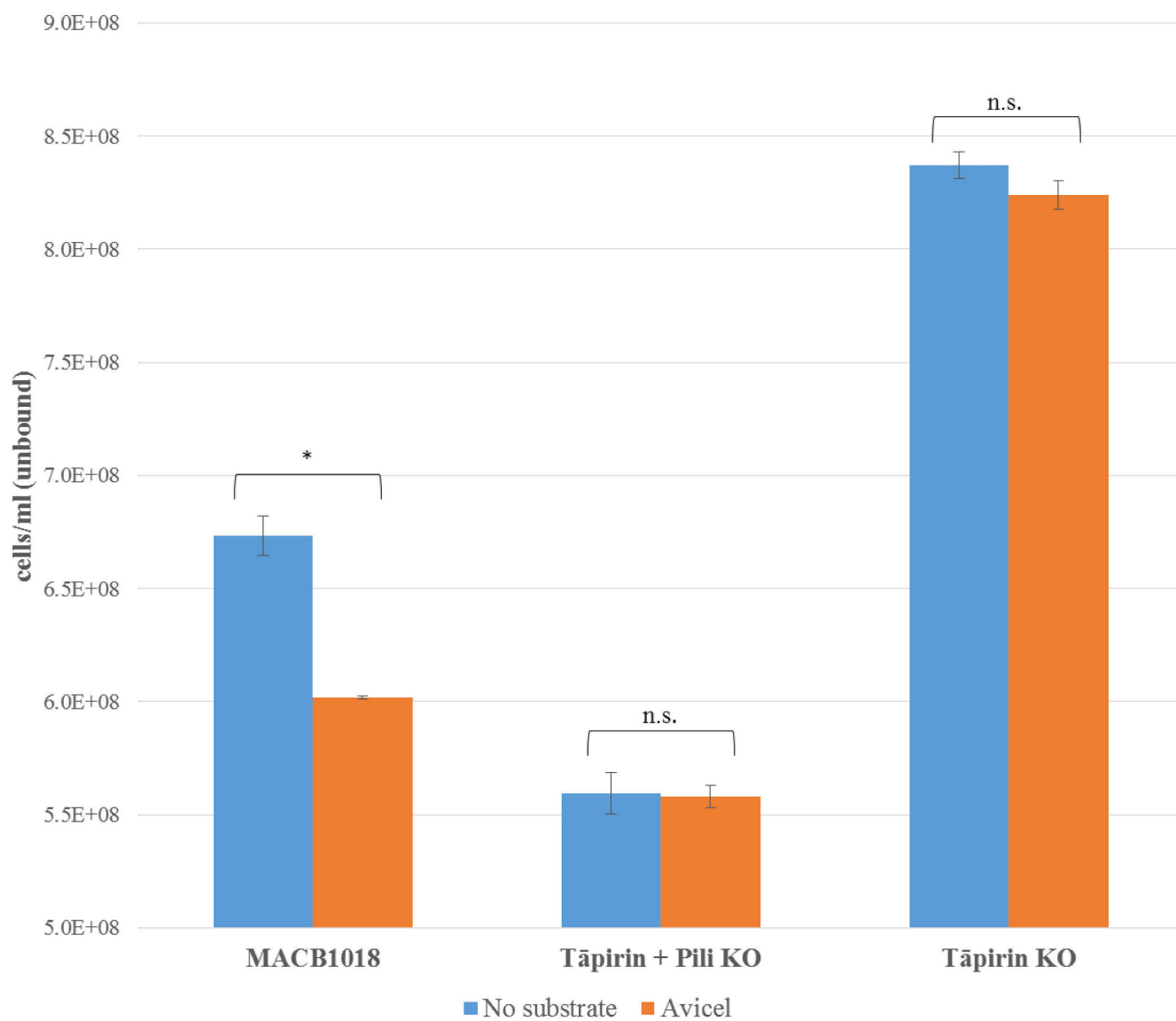
664

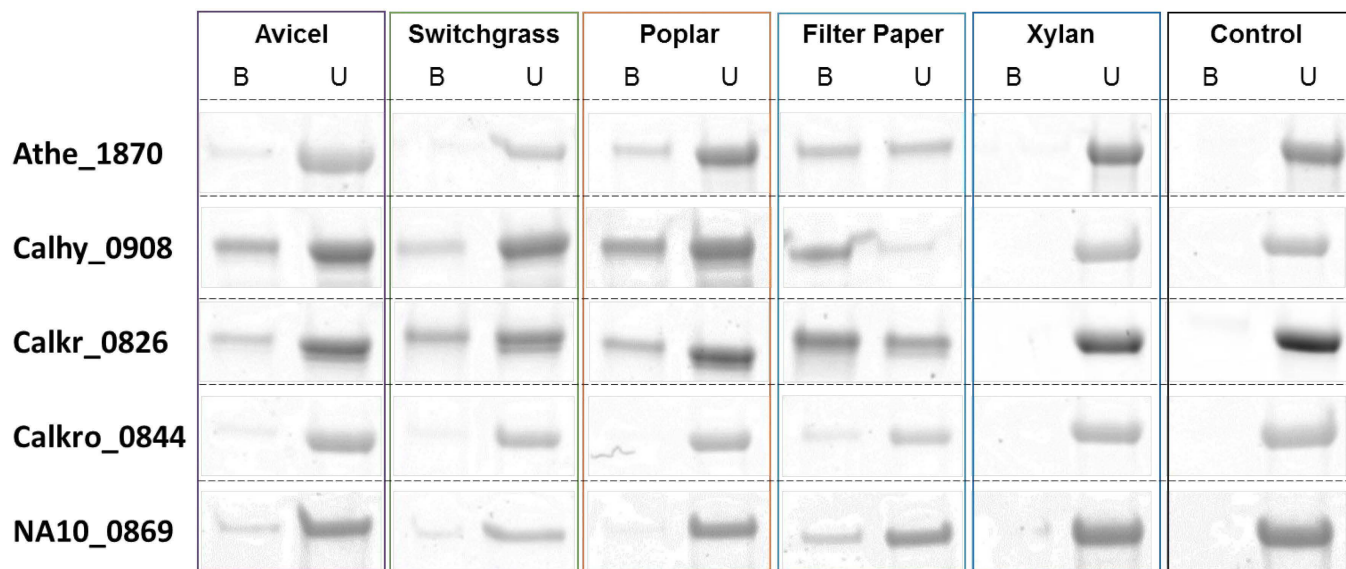
665 **Figure 7. Fluorescence microscopy of *Caldicellulosiruptor kronoyskyensis* using tāpirin**  
666 **antibodies.** *C. kronotskyensis* cells (in orange) grown on filter paper were incubated with  
667 primary antibody targeting the Calkro\_0844 tāpirin, and visualized (in green) with fluorescence  
668 microscopy and acridine orange staining (as done in (18)).

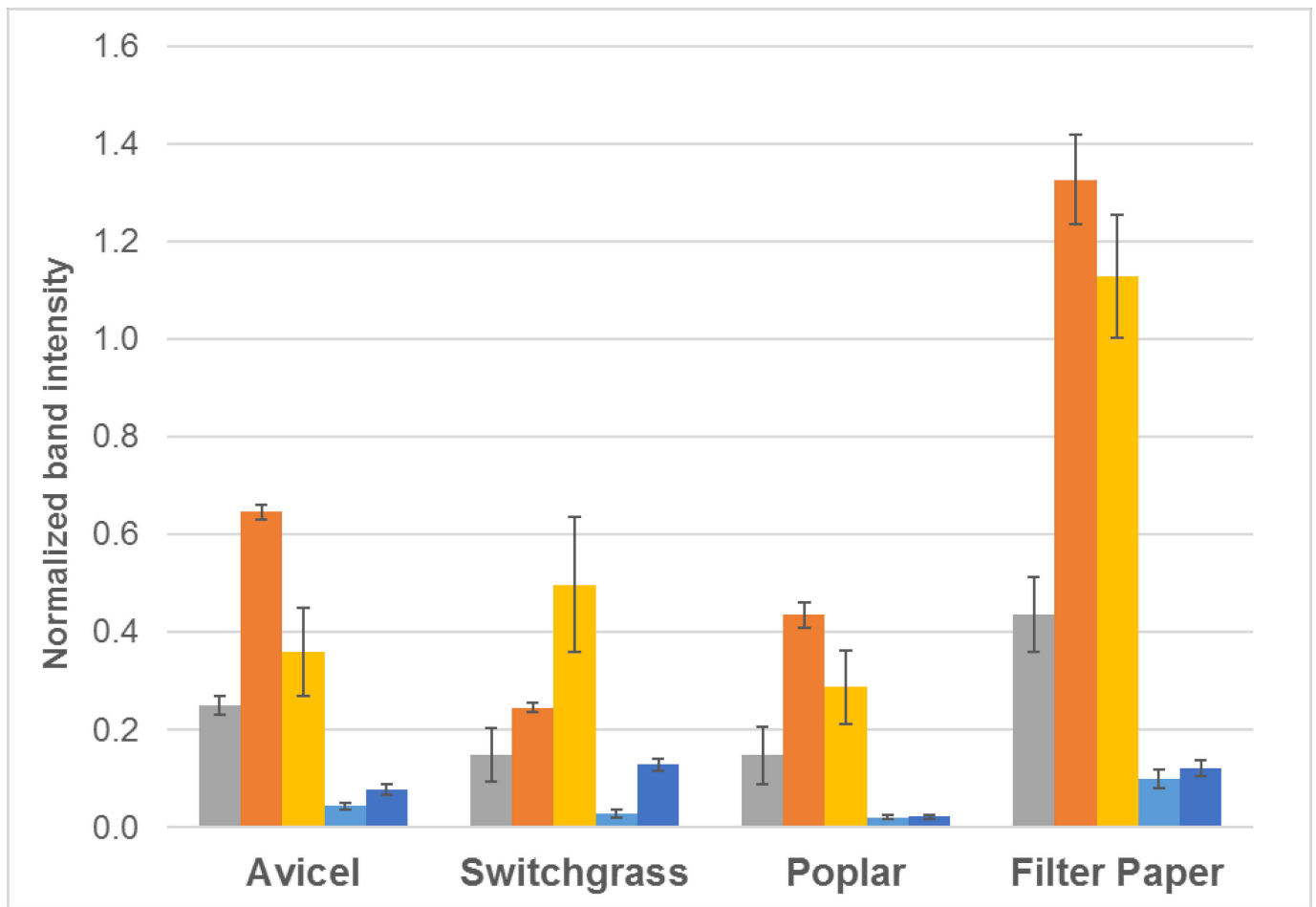
669

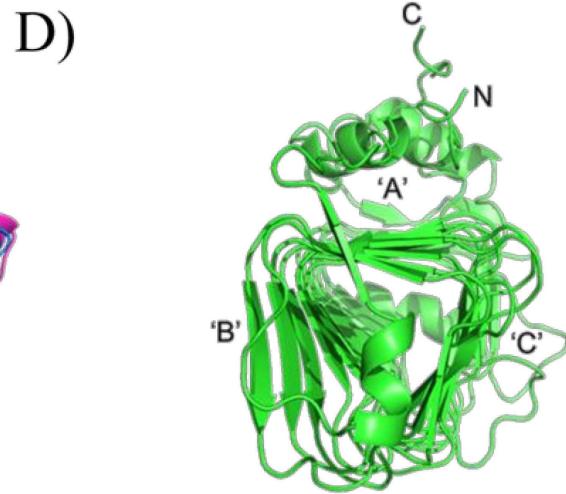
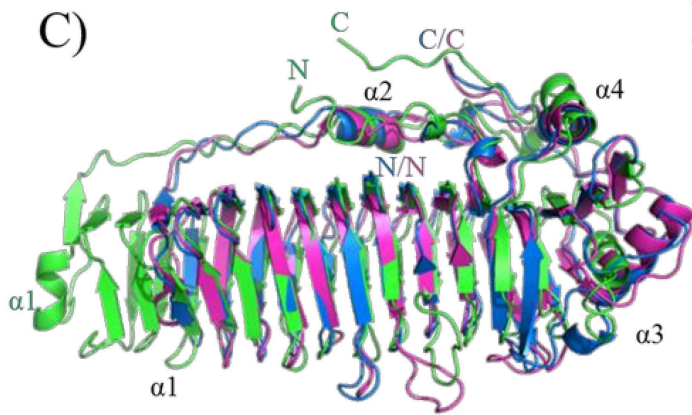
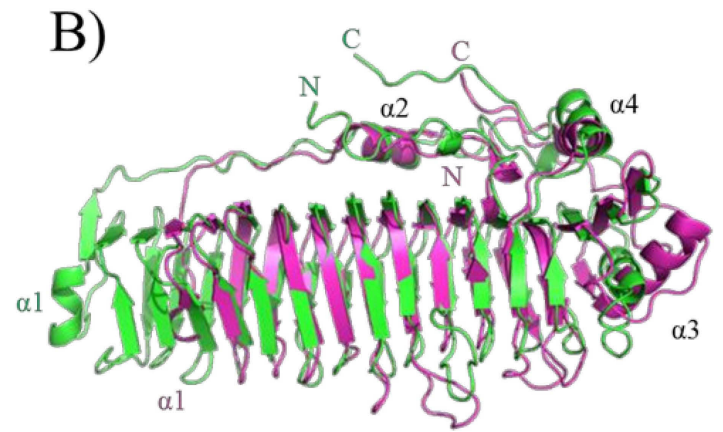
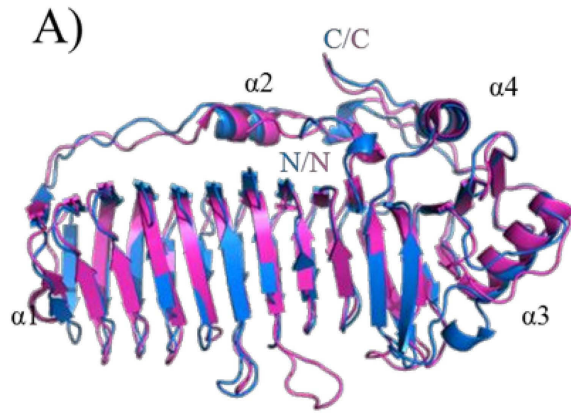
Highly to Weakly Cellulolytic Species

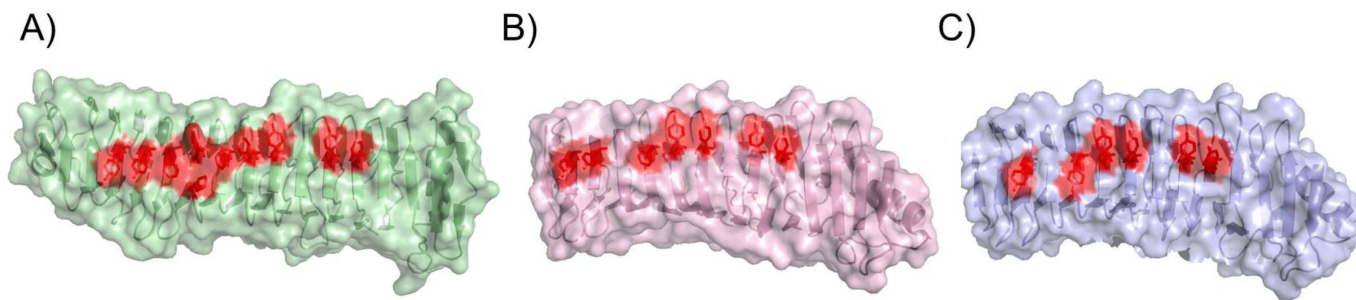
	Type IV Pili	Tāpirins		Glucan Degradation Locus	
		1	2		
<b>Calhy</b>	0896-0906	0908		Not Present	0 Glycoside hydrolases
<b>Calkr</b>	0812-0825	0826	0827	0017, 0086-0867 1848-1849, 2455, 2522	2 Glycoside hydrolases
<b>Athe</b>	1885-1872	1871	1870	1867-1844	6 Glycoside hydrolases
<b>Calkro</b>	0830-0843	0844	0845	0850-0864	6 Glycoside hydrolases
<b>NA10</b>	0883-0870	0869	0867 0864	0860-0848	7 Glycoside hydrolases

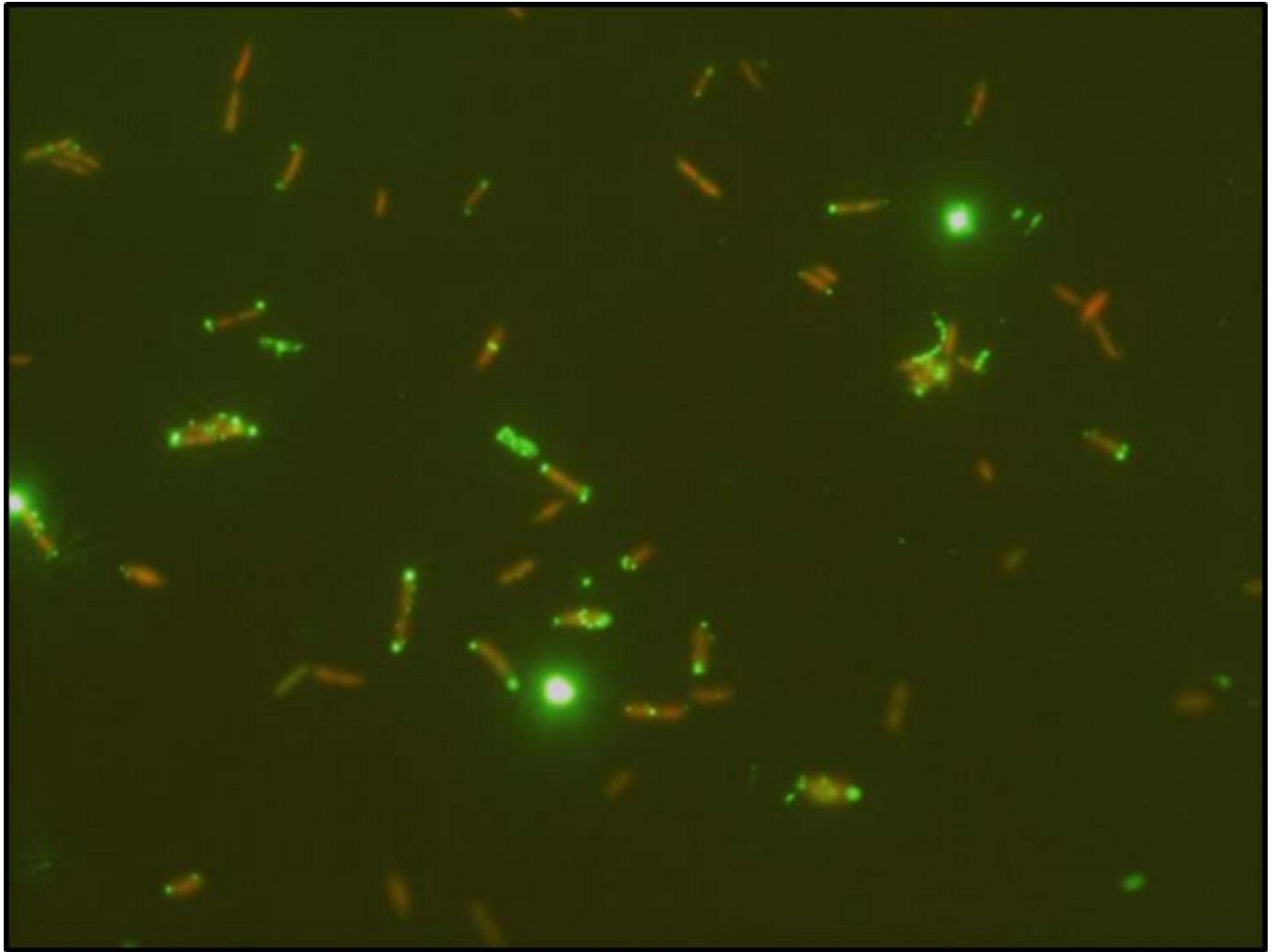












**Table 1. Characteristics of selected tāpirins from *Caldicellulosiruptor* species**

Species	Isolation Location	Tāpirin Gene Loci	# AA	M <sub>r</sub> (kDa)	%ID					%ID Scale
					Athe_1870	Calhy_0908	Calkr_0826	Calkro_0844	NA10_0869	
<i>C. bescii</i>	Kamchatka, Russia	Athe_1870	649	70.2	-	16.1	34.8	53.2	55.4	0
<i>C. hydrothermalis</i>	Kamchatka, Russia	Calhy_0908	638	70.8	16.1	-	16.7	15.9	17.3	25
<i>C. kristjanssonii</i>	Hveragerði, Iceland	Calkr_0826	634	68.8	34.8	16.7	-	35.7	34.8	50
<i>C. kronotskyensis</i>	Kamchatka, Russia	Calkro_0844	642	69.5	53.2	15.9	35.7	-	84.9	75
<i>C. naganoensis</i>	Nagano Prefecture, Japan	NA10_0869	642	69.9	55.4	17.3	34.8	84.9	-	100

**Table 2. Primers used in this study.** Underlined portions are specific to vector backbone or overlap between fragments. GA, Gibson Assembly construction; KLD, KLD reaction construction.

Primer	Sequence	Use
Athe_1885 5' fwd	<u>CGTTAAGGGATTTGGCTCTTTATCTAAAAGATGTTTTGTATC</u>	pLLL012 GA.
Athe_1885 5' rev	<u>CTAAAACAACCTGTATTGGATACAGTCTGATACACCTACCCCTCTATTTG</u>	pLLL012 GA
Athe_1870 3' fwd	<u>CAAATAGAGGGGTAGGTGTATCAGACTGTATCCAATACAGGTTGTTTTAG</u>	pLLL012 GA.
Athe_1870 3' rev	<u>GAAGATCCTTTTGATAATCTCATTAAGTAATTTTCTAACATTCTTAACCTAATTC</u>	pLLL012 GA
pGL100 fwd	<u>GAATTAAGTTAAGAATGTTAGAAAATTACTTAATGAGATTATCAAAAAGGATCTTC</u>	pLLL012 GA.
pGL100 rev	<u>CAAAACATCTTTTAGATAAAGAGCCAAAATCCCTTAACGTGAGTTTTTC</u>	pLLL012 GA
pGL100 fwd	<u>ATGAATAAAGATGCTTACATTCAAATG</u>	pLLL023 KLD
pGL100 rev	<u>TCTAGAGACCATCCTTTCTATGTAG</u>	pLLL023 KLD
Athe_1871 5' fwd	<u>CGTTAAGGGATTTGGCCTCCTTTTACCCCTCAC</u>	pLLL024 GA
Athe_1871 5' rev	<u>CCTCTTTTAAATCCTGTCCAGAAAACAAGTTAATGGTAAATGC</u>	pLLL024 GA
Athe_1870 3' fwd	<u>CAAAAAGAATTCCATTTTAAACAGGTTGTTTTAGTTTATTTACCTGATATTCC</u>	pLLL024 GA
Athe_1870 3' rev	<u>GATAATCTCATCTTTCTACACATGCTCTGCTCATC</u>	pLLL024 GA
P <sub>slp</sub> HTK fwd	<u>GCATTTACCATTAACCTTGTTTTCTGACAGGATTTAAAAGAGGCTATGC</u>	pLLL024 GA
P <sub>slp</sub> HTK rev	<u>CAGGTAATAAACTAAAACAACCTGTTAAAATGGAATTCCTTTTGAACATCAAC</u>	pLLL024 GA
pLLL023 fwd	<u>AGCAGAGCATGTGTAGAAAGATGAGATTATCAAAAAGGATCTTCACC</u>	pLLL024 GA
pLLL023 rev	<u>GGTAAAAGGAGGCCAAAATCCCTTAACGTGAG</u>	pLLL024 GA
Athe_1870 3'	CAAAATGAGCGTTTCCACTC / CCAACTAACAACCCAATAGTC	PCR screening
Athe_1885 5'	GAGGGGTACAGGTGATTG / CTATCATGTTCCGCTCATAGC	PCR screening
Athe_1871 5'	ACACACAATTCTTGATAAACAACC	PCR screening
HTK rev	<u>CCAATTTATGATCCAGGTGGTT / CCAATTTATGATCCAGGTGGTT</u>	PCR screening
Apr fwd	<u>CCCAAGGTTGAGAAGCTGAC</u>	PCR screening